

Open Research Online

The Open University's repository of research publications
and other research outputs

Data exploration and knowledge extraction: their application to the study of endocrine disrupting chemicals

Thesis

How to cite:

Roncaglioni, Alessandro (2008). Data exploration and knowledge extraction: their application to the study of endocrine disrupting chemicals. PhD thesis The Open University.

For guidance on citations see [FAQs](#).

© 2008 The Author



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Version: Version of Record

Link(s) to article on publisher's website:

<http://dx.doi.org/doi:10.21954/ou.ro.0000eb07>

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

UNRESTRICTED

Enrica
The Open University, UK

— Advanced School of Pharmacology —
Dean, Enrico Garattini M D

Mario Negri Institute for
Pharmacological Research

15/10/2008

DATA EXPLORATION AND KNOWLEDGE EXTRACTION: THEIR APPLICATION TO THE STUDY OF ENDOCRINE DISRUPTING CHEMICALS

Thesis submitted by

Alessandra Roncaglioni

Istituto di Ricerche Farmacologiche "Mario Negri", Milano, Italy

for the degree of

Doctor of Philosophy

Submission date: 31 March 2008
Date of award: 10 October 2008

MARCH 2008

ABSTRACT

Interest in computer-aided methods for investigating the biological field has increased significantly. One method is Quantitative Structure-Activity Relationships (QSAR), a valuable technique for predicting the effects of a substance from its chemical structure. A challenging application of QSAR is in characterizing the (bio)activity profiles of chemicals. Endocrine disrupters (EDs) are exogenous substances interfering with the function of the endocrine system and represent an interesting field of application for *in silico* methods. EDs targets include nuclear receptors, particularly effects mediated by the oestrogen receptor (ER).

They are also mentioned as substances requiring a more detailed control and specific authorisation within REACH, the new European legislation on chemicals. QSAR represents a challenging method to approach data gap about EDs since REACH substantially boosted interest on computational chemistry to replace experimental testing.

This work aimed to explore the status, availability and reliability of non-testing methods applied to endocrine disruption via oestrogen receptors and eventually to propose new models easily exploitable in regulatory contexts.

The work evaluated existing QSAR models present in literature to assess their validity on the basis of the OECD principles for QSAR validation. Different kinds of models have been analysed and they were externally validated with new data found in the literature.

Furthermore, new QSAR binary classifiers have been developed using different data mining techniques (e.g.: classification trees, fuzzy logic, neural networks) based on a very large and heterogeneous dataset of chemical compounds. The focus was given to both binding (RBA) and transcriptional activity (RA) better to

characterize the effects of EDs. A possible combination of the models was also explored. A very good accuracy was achieved for both RA and RBA (>85%). These models can be a valuable complement to *in vivo* and *in vitro* studies in the toxicological characterisation of chemical compounds.

ACKNOWLEDGMENTS

There are many persons I wish to thank for their essential support.

First of all my gratitude goes to the "Mario Negri" Institute for giving me the opportunity to undertake my PhD studies. In particular the head of my lab and my director of studies Emilio Benfenati is gratefully acknowledge for the important role he played while performing my research.

I'm really grateful to Prof. Werner Dubitzky for the fundamental advices he provided me, his helpfulness and also for his kind hospitality in Northern Ireland.

I have to acknowledge all my present and former colleagues in the lab since they have been often crucial to stimulate myself in doing a better job. I would like especially to mention Chiara for fruitful cooperation, Elena and Morena for sharing the beginning of our research experience, Antonio for revising my thesis draft, Mosè for tolerating my loud way of performing research, Silvia Rudy and Vittorio for the relaxing moments of break.

I'm grateful to Tanya Netzeva and Andrew Worth from ECB (Ispra, Italy) for their support in the validation exercise and to all the colleagues I met through the EC projects I participated for the stimulating scientific environment I could experience with all of them.

There are no words to express how I appreciated the assistance of my friends at BCX (Orléans, France) for the use of their proprietary software and particularly Marco and Nadège for my memorable stay in France.

I'm thankful to my family, especially to my parents for always believing in me and to all my friends for their support.

Lastly, not least, I'm grateful to Andrea for having shared with me so many experiences of our lives, including this important target.

TABLE OF CONTENTS

Preface	1
 PART I: INTRODUCTION ON QSAR AND ENDOCRINE DISRUPTERS	6
CHAPTER 1 The endocrine disrupters issue	7
1.1. The oestrogen receptor (ER)	8
CHAPTER 2 QSAR and other computational modelling methods.....	13
2.1. <i>In silico</i> tools as an alternative to animal testing	13
2.2. Computational modelling methods	16
2.2.1. (Q)SAR	17
2.2.1.1. Models on ER with (Q)SAR	24
2.2.2. 3D-QSAR.....	26
2.2.2.1. Models on ER with 3D-QSAR.....	29
2.2.3. Virtual docking	31
2.2.3.1. Models on ER with docking	34
2.3. Conclusions	36
 PART II: VALIDATION OF EXISTING MODELS	39
CHAPTER 3 Critical assessment and selection of existing models for EDs	
accordingly to the OECD principles	40
3.1. Criteria for model selection.....	41
3.1.1. Preliminary evaluations of the OECD principles.....	42
3.1.2. Definition of the criteria suitable for short-listing of QSAR	44
3.2. Selected models	46

CHAPTER 4 Independent validation of selected (Q)SARs on the basis of the OECD principles	48
4.1. Materials and methods.....	48
4.1.1. Selected models to validate	48
4.1.1.1. Sources for experimental activity data of the models	48
4.1.1.2. Descriptors	49
4.1.1.3. Mathematical formulation of the models	50
4.1.2. Internal validation and model reproducibility	51
4.1.2.1. Re-calculation of the models and statistical assessment	52
4.1.2.2. Sensitivity study	53
4.1.3. External validation procedures	54
4.1.4. Applicability domain assessment	59
4.2. Results and Discussion	60
4.2.1. Model 1	60
4.2.1.1. Internal validation	60
4.2.1.2. Sensitivity study	60
4.2.1.3. External validation with new test sets.....	62
4.2.1.4. Applicability domain analysis	64
4.2.2. Model 2	67
4.2.2.1. Internal validation	67
4.2.2.2. External validation with a new test set	68
4.2.2.3. Applicability domain analysis	69
4.2.3. Model 3	70
4.2.3.1. Internal validation	70
4.2.3.2. External validation with a new test set	70
4.2.3.3. Applicability domain analysis	71
4.3. Conclusions	73

PART III: DEVELOPMENT OF NEW QSARs FOR OESTROGENICITY	74
CHAPTER 5 Binary classification models for screening heterogeneous datasets for their oestrogenic activity.....	75
5.1. Introduction.....	75
5.2. Material and methods	76
5.2.1. Dataset.....	76
5.2.2. Modelling methods.....	78
5.2.2.1. Classification and Regression Tree (CART)	78
5.2.2.2. Decision Forest (DF)	81
5.2.2.3. Adaptive Fuzzy Partition (AFP)	81
5.2.2.4. Multilayer perceptron (MLP)	86
5.2.2.5. Support Vector Machine (SVM)	89
5.2.3. Performance evaluation	91
5.3. Results and discussion	92
5.4. Conclusions	102
CHAPTER 6 Conclusions	104
References	109
List of abbreviations	125
Annex A	128
Annex B	157

LIST OF FIGURES

Figure 1.1 Structural representation of the ER. Picture modified from NURSA: The Nuclear Receptor Signaling Atlas © 2003, (www.nursa.org) [14].....	9
Figure 1.2 Examples of chemical groups known to bind to ER: natural hormones (1), synthetic hormones (2) and antioestrogens (3), phytoestrogens (4, 5) and xenoestrogens (6, 7). S indicates the Tanimoto similarity coefficient calculated using E ₂ as reference in ACD/Labs 9.0.	10
Figure 2.1 Schematic representation of the steps for developing (Q)SAR models.....	18
Figure 2.2 Representation of the lattice used to calculate field-based descriptors. 3D-QSAR descriptors are calculated by placing appropriately overlapped molecules within a lattice and calculating with a probe (placed in the upper left hand corner in the figure) the steric and electrostatic field contributions of the molecule at that point. Each matrix column contains the contribution to the field for each molecule at a specific grid point.	27
Figure 2.3 The process of mutual recognition between a protein (P) and a ligand (L) is governed by the Gibbs free energy of binding, released when ligand and receptor bind.....	32
Figure 4.1 The flowchart shows the rules for the SAR model as reported in the DSSTox database [111].....	51
Figure 4.2 Correlation of the activity data found in the NCTRER database, used to develop Model 1, and the METI database. In the graph – 5 value was arbitrarily assigned to inactive compounds. The "undefined" series contains 9 compounds which are inactive in the Japanese database but active in the NCTRER database (with activity values < -3.25), 36 compounds inactive in	

both databases and 10 compounds inactive in the NCTRER database but active in the Japanese database (with activity values < -1.15).....	56
Figure 4.3 Correlation for the values of "Energy of torsion" descriptor. Three compounds in red in the plot are outliers and were not considered to derive the correlation coefficient.....	61
Figure 4.4 Correlation for the values of "HOMO Energy" descriptor calculated with TSAR or CODESSA with AM1 parameterisation.....	61
Figure 4.5 Experimental versus calculated activity for the model equations using the original descriptor set or those recalculated with E _{HOMO} from alternatively TSAR or CODESSA.	62
Figure 4.6 Predicted activity (logRBA range) distribution for the test set of inactive chemicals of NCTRER dataset.	63
Figure 4.7 Comparison of the error extents for the training set and the two external test sets. Prediction performances are superior for EDKB test set.	64
Figure 4.8 PCA score scatter plot for the first two PC (explained variance: 59.4%), calculated on the basis of the eight descriptors, for the training set (active compounds, blue triangles) and the test set of inactive compounds (red triangles). Three outliers are indicated by red circles.	65
Figure 4.9 Percentage of compounds with residuals within one of two Log units for the first test set. Different ways for estimating applicability domain are compared with the reference values (pale blue and turquoise green dotted lines) where all compounds belonging to the test set are used.....	66
Figure 4.10 Percentage of compounds with residuals within one of two Log units for the second test set. Different ways for estimating applicability domain are compared with the reference values (pale blue and turquoise green dotted lines) where all compounds belonging to the test set are used.....	67
Figure 4.11 Comparison of the classification performances for the training and the test set.....	68

Figure 4.12 PCA score scatter plot for the training and test set using the first three components (explained variance = 47.2%).	69
Figure 4.13 PCA score scatter plot for the training and test set using the first three components (explained variance = 56.2 %).	71
Figure 4.14 Comparison of the classification performances for the training and the test set.	72
Figure 5.1 Compounds distribution in the classes of activity for RBA and RA.	77
Figure 5.2 Example of adaptive fuzzy partitioning of a bi-dimensional space.	84
Figure 5.3 A MLP neural network structure. Perceptrons are arranged in layers.	87
Figure 5.4 Exemplification of the SVM process for identifying the support vectors and the maximal margin classifier hyperplane. Adapted from [134].	89
Figure 5.5 ROC comparison of the RBA models obtained with different algorithms.	93
Figure 5.6 CART tree for RBA endpoint.	96
Figure 5.7 ROC comparison of the RA models obtained with different algorithms.	98
Figure 5.8 Comparison of misclassified FN for single and combined RBA models.	101
Figure 5.9 Comparison of misclassified FN for single and combined RA models.	101

LIST OF TABLES

Table 2.1 Overview of the methods that can be adopted to study the biological effect of chemical compounds.	37
Table 3.1 Models selected for external validation.	47
Table 4.1 List of descriptors used in <i>Model 1</i> and explanation of the procedure used to recalculate them.	54
Table 4.2 Comparison of the activity class assigned to compounds in common in the NTP compilation, as reported by Sutherland <i>et al.</i> [74], and NCTRRER dataset.	58
Table 4.3 Descriptor range and outlier descriptor values.	64
Table 4.4 Applicability domain evaluation with AMBIT Disclosure.	66
Table 4.5 Classification performances for the compounds belonging to the test set considered in or outside the AD.	73
Table 5.1 Compounds repartition in training validation and test sets.	78
Table 5.2 Parameter setting used for GA and MLP.	88
Table 5.3 Parameter setting used for GA and SVM.	91
Table 5.4 Overview of RBA results.	93
Table 5.5 List of selected variables in the RBA models.	94
Table 5.6 Misclassification rates for the CART terminal nodes for RBA. Node purity can be used to assess prediction reliability. Asterisks identify nodes with a misclassification rate greater than 0.3.	97
Table 5.7 Overview of RA results.	97
Table 5.8 List of selected variables in the RA models.	99
Table 5.9 Performances of the combined models for RBA and RA.	100

PREFACE

Interest in computer-aided methods for investigating biological events has increased significantly in recent years. Analogously to the expressions *in vivo* (referring to methods using animals) and *in vitro* (referring to methods using mainly cellular systems), the expression *in silico* has been introduced referring to silicon, as a metaphor for computers. *In silico* tools are becoming more accessible to researchers as their cost drops and the speed of computational calculation increases, so interest in their application can spread to a wide range of biological problems. Data mining techniques are frequently used to analyse biological data such as genomic or proteomic findings, for example. A challenging application of these methods is modelling and characterizing the (bio)activity profiles of chemicals. Many studies have addressed, for example, ecotoxicity or human health. These methods seek relations between chemical structures and the observed properties exhibited by the compounds under investigation. Among the properties that can be analysed *in silico*, endocrine disruption forms an emerging field that is attracting attention from scientists and political institutions. Endocrine disruptors (EDs) represent a number of exogenous substances interfering with the function of the endocrine system, producing consequences on the homeostasis of all the processes controlled by this system in humans and wildlife. The EDs issue is highly complex on account of the wide range of mechanisms of action they can interfere with. The targets include receptors belonging to the nuclear receptor superfamily. Among them the receptor more extensively investigated to account for endocrine disrupting effects is the oestrogen receptor (ER) mediating the effects of the steroid hormone 17 β -estradiol. In this field *in silico* techniques

could be a valuable complement to *in vivo* and *in vitro* studies in assessing hazard of chemicals.

Many computational methods can be used to analyse the toxicity or biological activity of chemicals, particularly as regards their interactions with biological macromolecules such as receptors.

QSAR and other *in silico* tools are very suitable for addressing the direct interaction of chemicals with receptors, since both the ligands and the receptors can be characterised by their chemical structure. The ER was one of the first targets studied with computer-aided methods addressing specifically EDs especially because of the great number of data available to be modelled.

One of the eligible methods is Quantitative Structure-Activity Relationships (QSAR), a widespread and valuable technique for predicting the risk of a substance from its chemical structure. It is based on the assumption that a relationship exists between molecular features encoded in chemical descriptors and the biological activity or biochemical property. 3D-QSAR relies on field-based descriptors derived by mapping the environment surrounding the molecules in terms of energetic interactions of various natures once the molecules are properly aligned in the space. This is obtained by placing the molecule within a lattice and calculating the interaction energies of that molecule with a probe at each point of the grid.

Virtual docking is a methodology to predict computationally the binding between two molecules, usually a protein and a small molecule (ligand).

It is difficult to compare the results of different approaches since often these models rely on different datasets or at least on different validation procedures to assess their reliability. A further aspect to be considered is the intended use of the model.

A very challenging field of application of these methods is predictive toxicology, particularly to satisfy regulatory requirements. The main obstacle to completing

hazard assessment of chemicals is the lack of adequate experimental data, required to cover all the major effects relating to human health or ecological safety.

In the near future this situation will change with REACH (Registration, Evaluation, Authorisation and restriction of Chemicals), the new European legislation on chemicals. REACH legislation has substantially boosted interest in computational chemistry to replace experimental testing since it provides some indications for the use of existing information, techniques such as QSARs, read-across and analogue identification, to avoid unnecessary testing. Some studies have estimated that these alternatives, including QSARs, will reduce the additional costs due to implementing REACH by about one billion Euros and these alternatives could potentially save more than a million animals.

Some efforts are still needed to facilitate regulatory acceptance of QSAR as an alternative by increasing the transparency and reproducibility of the models generated with QSAR and by taking account of regulators' needs. The OECD has identified some principles of QSAR validation to satisfy these aspects.

In particular EDs are mentioned as substances requiring more detailed control and specific authorisation within the REACH framework together with other groups of chemicals of particular concern.

Overall on the one hand there is the need to cover data gaps, but on the other hand no technique is yet able to deal efficiently with this need; computational toxicology can be of help in this task but more efforts to gain a wider acceptance are still required. The main issue this work aims to contribute is to provide an insight about the use of *in silico* models to address EDs effects mediated through the oestrogen receptor.

In particular this work aims to explore the status, availability and reliability of non-testing methods applied to endocrine disruption mediated through the oestrogen

receptor and eventually to propose new models easily exploitable in regulatory contexts.

Therefore the project objectives and their implementation adopted during this work are:

- To analyse the existing information and models in order to: i) assess what level of accuracy was already obtained and ii) identify the most promising methods to be explored further. This was done by close investigation of the literature to identify existing *in silico* models developed for EDs with QSAR, 3D-QSAR or docking approaches.

Secondly, the focus was given to QSAR only, the method more frequently used and better characterised for its use in regulatory assessment, in order to validate promising models. Complementarily, this process involved collecting experimental data, either to validate the models identified in literature, or to use them for developing new models.

- To validate existing models. This included both internal and external validation. Internal validation aims to analyze model reproducibility and transparency of all steps including descriptor calculation and the application of the algorithm suggested in the literature. External validation is focused on the assessment of performances in prediction by using newly identified compounds and on the assessment of applicability domain.
- To apply different modelling techniques, studying new models for oestrogenicity to be of practical use as fast and reliable screening method, by adopting the most promising approaches identified during the previous steps of the project.

This thesis is divided in three conceptual Parts for a total of six Chapters. In the first introductory Part, the general issue of EDs and oestrogen-mediated effects, in particular, is presented (Chapter 1). This is followed by an introduction on *in silico*

methods to study the interaction of chemicals with receptors including a review of existing models for EDs effects mediated through the ER (Chapter 2).

In the second Part the literature findings have been rationalised and key factors for the practical application of QSAR in the context of regulatory framework are discussed. This includes some theoretical considerations on desirable characteristics of QSAR and a detailed discussion on OECD principles for QSAR validation from a practical point of view (Chapter 3). Chapter 4 reports the validation exercise of the three most promising models identified in the literature.

The third Part describes in Chapter 5 the development of new binary classification models for addressing oestrogenic effects (binding and transcriptional activity of ER). Finally the sixth Chapter draws the main conclusions on the overall themes developed in the thesis.

PART I

INTRODUCTION ON QSAR AND ENDOCRINE DISRUPTERS

CHAPTER 1

THE ENDOCRINE DISRUPTERS ISSUE

The issue of chemical compounds interfering with the endocrine system, commonly called endocrine disrupters (EDs), is an emerging field of high concern for scientists [1] and political institutions [2,3].

An endocrine disrupter is an exogenous substance or mixture that alters functions of the endocrine system and consequently causes adverse health effects in an intact organism, its progeny, or populations in both humans and wildlife. As a consequence effects on reproductive, developmental, immunological and neurological functions may occur such as cancer, behavioural changes and reproductive abnormalities [4].

A great number of targets can be affected by EDs and thus a wide range of substances are suspected to be endocrine disrupters [5]: they range from pesticides to plastics additives such as phthalates, from environmental pollutants (like PCBs, dioxins and furans) to flame retardants, from cosmetic ingredients to natural products. Many of these chemicals, such as plasticizers, are relevant in industrial processes and some of them are Persistent Organic Pollutants (POPs). Due to the key role of many of these substances in industrial processes, it is essential to fill the data gap also in view of new legislation requirements.

The high complexity of the EDs issue is related to the wide range of mechanisms of action they can interfere with, since they can directly damage an endocrine organ, directly alter the function of an endocrine organ, interact with receptors or alter hormone metabolism, inhibiting steroidogenesis or increasing hepatic metabolism and clearance [6].

In particular, amongst the receptor targets a series of receptors belonging to the Nuclear Receptor (NR) superfamily can be enumerated. This is a group of ligand-inducible transcription factors that mediate the effects of hormones and other endogenous ligands to regulate the expression of specific genes [7]. Among them are included receptors for various hormones like steroids, retinoic acid and thyroid hormones. This family also includes some "orphan" receptors whose natural substrates are still unknown [8].

QSAR and other *in silico* tools work very well in addressing the direct interaction of chemicals with receptors since either the ligands or the receptors can be characterised by their chemical structure.

In this project attention has been focused on the study of chemicals interfering with the oestrogen receptor (ER) and more details about this specific system will be introduced in the next paragraph.

1.1. THE OESTROGEN RECEPTOR (ER)

The oestrogen receptor is a ligand-activated transcription factor that mediates the effects of the steroid hormone 17 β -estradiol (E2) in both males and females [9]. It is involved in the development, growth and maintenance of reproductive tissues but it is also present in a number of non-reproductive tissues such as bone, liver, brain, the CNS, cardiovascular and immune systems in the physiological situation. ER seems also to be involved in pathological processes such as osteoporosis or breast cancer [10,11].

ER is present in two isoforms (ER α and ER β) in humans that exhibit overlapping but distinct tissue distribution patterns and differ in their ligand-binding ability and transactivational properties [12,13].

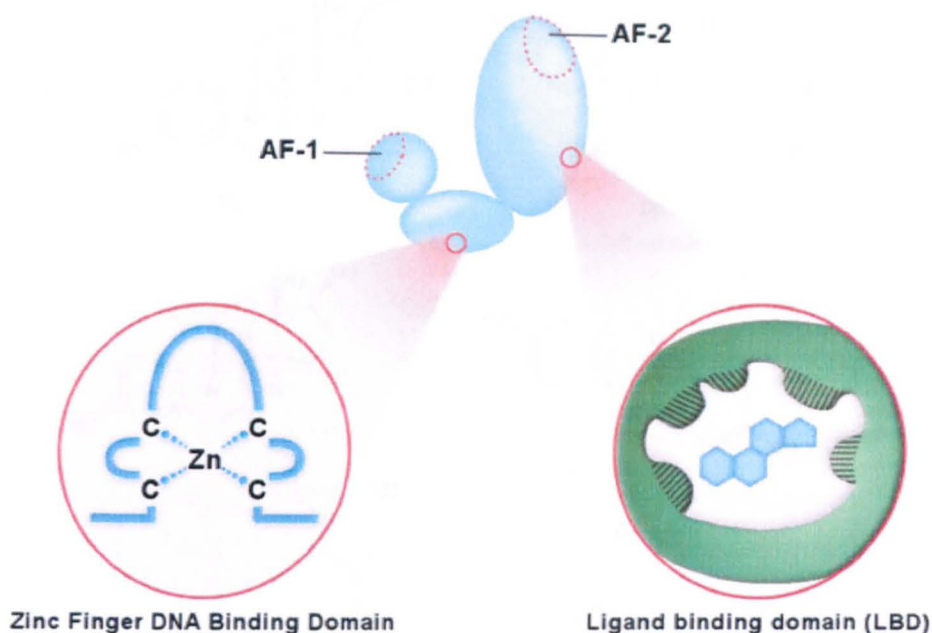


Figure 1.1 Structural representation of the ER. Picture modified from NURSA: The Nuclear Receptor Signaling Atlas © 2003, (www.nursa.org) [14].

Three functional regions shown in Figure 1.1 can be identified in the ER structure [15]:

- The Ligand binding domain (LBD) [16]

The LBD constitutes the COOH-terminal region of ER and mediates ligand binding to the endogenous hormone E2. The binding cavity of human ER α has an accessible volume of 450 Å³, quite large compared to the natural ligand (250 Å³) [17]. Other chemicals that share a relatively low similarity with E2 can bind with different degrees of affinity and sometimes with a diverse specificity to ER subtypes. Figure 1.2 reports some examples of different categories of substances known to bind ER. Beside the natural hormone (1) other chemicals binding in the LBD of ER include some drug-like synthetic hormones (2), antioestrogens (3), oestrogens produced by plants (4) or fungi (5) and chemical contaminants like PCBs (6) and alkylphenols (7).

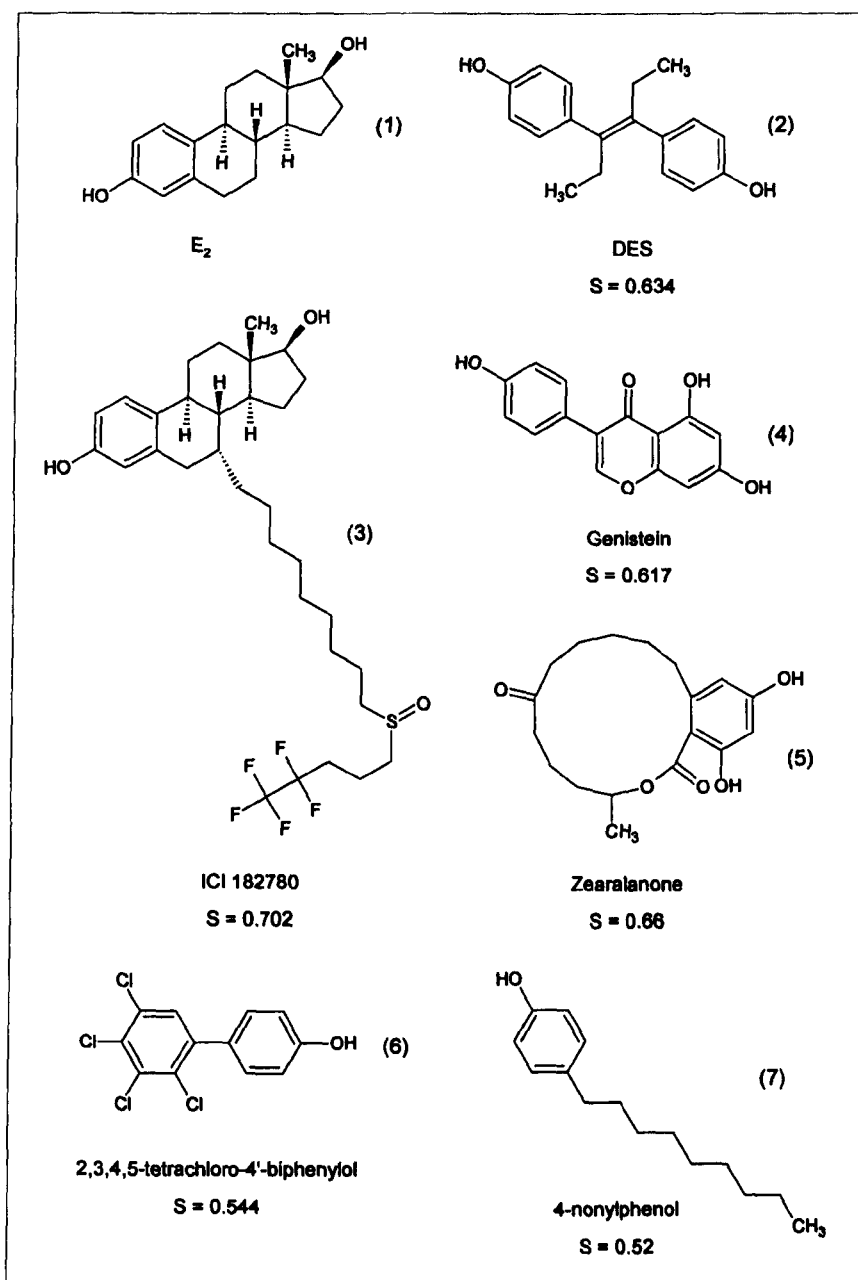


Figure 1.2 Examples of chemical groups known to bind to ER: natural hormones (1), synthetic hormones (2) and antiestrogens (3), phytoestrogens (4, 5) and xenoestrogens (6, 7). S indicates the Tanimoto similarity coefficient¹ calculated using E_2 as reference in ACD/Labs 9.0.

¹ The Tanimoto coefficient is calculated with the following formula: $S = N_c / (N_q + N_t - N_c)$ where N_q , N_t and N_c are respectively the number of bit screens set in the query structure, in the target structure, and those in common to both structures.

The LDB also contains an Activation Function (AF-2), whose structure and function are governed by the binding of ligands. Crystallographic investigation of the LBD has shown that, when ER is bound to an agonist, this region is responsible for recruiting coactivator peptides and starting the cascade of events that provoke the transcription of oestrogen-regulated genes. This process includes receptor dimerisation, receptor-DNA interaction, interaction with coactivators and other transcription factors, and formation of a preinitiation complex [18]. On the other hand, binding to compounds with antagonistic activity causes a conformational change in the position of Helix-12 that occupies the space devoted to the interactions with coactivators blocking the transcription process [19].

- The DNA binding domain (DBD)

DBD contains two zinc ions, each co-ordinated by the tetrahedral arrangement of thiol groups from four cysteine residues, in a zinc finger structure. This area plays an important role in receptor dimerisation and in binding of receptors to specific DNA sequences called estrogen response elements (EREs) [20]. ER α and ER β can also modulate the expression of genes without directly binding to DNA.

- The NH2-terminal region contains a ligand-independent activation function (AF-1) that is again involved in the interaction with transcription factors [18].

Since the ER is involved in pathological processes, the receptor, the ligands, and all relevant accessory proteins have often been considered as targets in pharmaceutical research to develop therapeutic drugs for hormone-related diseases. For this reason the ER was investigated in depth with a series of *in silico* techniques in order to facilitate the drug discovery process. In this kind of study

usually a small, quite homogeneous group of synthetic compounds is taken into account and the highest grade of mechanistic information is included. This implies that not only the chemical itself but also the receptor is considered in the analyses (e.g. virtual docking studies) and these studies are coupled with other methods like crystallography and site-directed mutagenesis techniques [21]. Even if these experiments are outside the purposes of the present project, the interest from the pharmacological point of view was the driving force to elucidate key aspects of the ER functioning, briefly introduced above, and to provide a mechanistic insight on the receptor-ligand binding and the agonist/antagonist interactions.

Today the interest in nuclear receptors is no longer limited to the drug discovery process but includes also the endocrine disrupters issue. A more detailed analysis of the modelling studies already conducted in this field will be provided in Chapter 2.

To detect endocrine-disrupting effects, different *in vivo* and *in vitro* experiments have been set up. Since the system itself and the variety of targets is very large and variable among different species, a battery of tests has been proposed to analyse different possible interactions with the endocrine system, organised in a tiered approach [22,23]. Among them, some *in vitro* tests have been designed to detect the direct binding with the LBD (receptor binding assay), or the transcriptional activation of DNA (cell proliferation and reporter gene assays) [24]. Despite there being many experimental methods, both *in vivo* and *in vitro*, for encoding information on the disruption activities of chemicals, only a few experimental protocols are fully standardised and validated, so computational chemistry can be a complementary tool to characterize properties of endocrine-disrupting chemicals.

CHAPTER 2

QSAR AND OTHER COMPUTATIONAL MODELLING METHODS

Nowadays the interest in employing computational modelling techniques to predict the activity of chemicals is constantly growing. Many computational methods can be used to analyse the toxicity or biological activity of chemical compounds, in particular in relation to their interaction with biological macromolecules (i.e. receptors) as well as for modelling other physico-chemical properties. An overview of these methods will be provided along with some examples of different techniques applied to the prediction of oestrogen receptor (ER)-mediated responses. The applications involve studies in drug design, for selecting lead compounds to develop new drugs, and others involving a wide range of substances, called endocrine disrupters, which can interfere with the mechanisms of hormone regulation as described in Chapter 1. Molecular modelling techniques such as Quantitative Structure-Activity Relationships (QSAR) and related methods like Comparative Molecular Field Analysis (CoMFA) and virtual docking have been used to investigate these phenomena and will be described here. Implications concerning the regulatory acceptance and use of these methods and the derived models for hazard identification and priority setting will be also addressed.

2.1. *IN SILICO* TOOLS AS AN ALTERNATIVE TO ANIMAL TESTING

In recent years interest in computer-aided methodologies and *in silico* techniques applied to the investigation of biological activities has significantly increased. *In silico* tools are becoming more and more accessible to researchers because of

their decreasing cost and increasing speed of computational calculation; for these reasons interest in their application is spreading to a wide range of biological problems. Frequently data-mining techniques are used to analyse a wide range of biological data like, for example, those from genomics [25,26] or proteomics [27,28].

Among them, one challenging application is the modelling and characterisation of the (bio)activity of chemical compounds. A variety of techniques is included in this area, and among them QSAR has been widely used in characterizing many properties of chemical compounds, using information derived from chemical structures. Many studies have been published addressing for example ecotoxicity [29,30], human health [31-33], physico-chemical [34,35] and Absorption Distribution Metabolism Excretion (ADME) properties [36,37].

Nowadays an (eco)toxicological characterisation of the risk associated with the use of industrial chemicals is still incomplete. Two elements are required to address this issue: an evaluation of the intrinsic properties of the chemical - hazard assessment, and an estimation of the exposure. The primary constraint in completing the hazard assessment is the lack of adequate experimental data required to cover all relevant effects of chemicals associated with human health or ecological safety. A study on the availability of the toxicity data for High Production Volume (HPV) existing substances revealed significant gaps in publicly available knowledge about these chemicals [38,39].

In the near future the situation is going to change since new requirements have been introduced by the legislation recently approved in the EU called REACH [40]. REACH (Registration, Evaluation, Authorisation and restriction of CHemicals) will modify the current authorisation scheme for commercialising chemical products within the EU market. Previous legislation distinguished between "existing" and "new" chemicals (i.e. those placed on the market after 1981); while "new" chemicals have to be tested before they are placed on the market,

there were no such provisions for "existing" chemicals. Thus, although some information exists on the properties and uses of existing substances, this system has not produced sufficient information about the effects of the majority of existing chemicals on human health and the environment. With REACH it is foreseen to fill this data gap but, as a consequence of the increased need of experimental tests, a number of problems have to be faced, ranging from economic - in terms of either time or intrinsic costs - to ethical ones. It is for example estimated that 10.7 million laboratory animals are sacrificed in testing each year in Europe, and 10% of these tests are for regulatory purposes [41]. The economic costs are also very high: over the next 11 years the estimated costs for testing the approximately 30,000 existing substances produced above 1 tonne/year would be nearly 2.5 billion € [42].

In silico techniques can play a key role in this process, being a valuable complement to *in vivo* and *in vitro* studies in assessing hazard of chemicals. Of course they cannot substitute for laboratory experiments, but they can be adequately integrated with them. In this way they can provide a prioritisation of compounds needing deeper *in vitro* and *in vivo* investigations, allowing a more rational use of resources by better planning of experimental testing.

General rules are set out in the REACH legislation for waiving of tests through the use of existing information, techniques such as (Q)SARs and read-across. Some studies have estimated that by implementing the use of these alternatives, including (Q)SARs, it is possible to reduce the additional costs due to the REACH implementation by about one billion € [43] and the use of these alternatives can potentially save 1.3 – 1.9 million animals [44].

Some efforts have still to be made to facilitate the regulatory acceptance of QSAR as an alternative approach [45] by increasing the transparency and reproducibility of the models generated with QSAR, and by taking into account regulators' needs.

Even though in view of REACH legislation the interest in computational chemistry to replace experimental testing has increased significantly, it should be noted that this is not the only case where QSAR is accepted in a regulatory framework. In the U.S.A., Canada, Japan and some of the EU countries, at a national level, these methods are already accepted and used [46,47].

In silico technology can also represent a valid instrument to improve the safety of newly synthesised chemical compounds or even chemicals in a pre-synthesis phase, since only the chemical composition of the compounds is required. This can help in evaluating safer alternatives to be put on the market.

2.2. COMPUTATIONAL MODELLING METHODS

In this section an overview of different computational modelling methods will be provided. Based on a literature search, all promising studies on ER interfering chemicals through the use of computational chemistry were reviewed. This part of the work served as a basis to identify pros and cons of the different approaches and to give background on what was already achieved in this area, and on the most promising approaches already adopted, the better to set the experimental scene.

Computational studies dealing with EDs have to some extent a different perspective and use generally slightly different tools from those in the pharmacological area. The main objective is to provide a tool for prioritising chemicals, so detailed information on the specific binding mode of different chemical categories is often not available. For the same reason the primary concern in screening environmental contaminants for their possible interference with endocrine systems is to identify all potential EDs: the target is to avoid the so-called false negatives. In contrast, pharmaceutical industries in their virtual HTS studies want to focus on the most likely active compounds, avoiding false positives.

The ER was one of the first targets studied with computer-aided methodologies specifically for EDs. The use of QSAR applied to endocrine disrupters became in recent years quite widespread [48-50] and the U.S. Environmental Protection Agency (U.S.-EPA) was involved in the endocrine disrupters screening programme [51], with the goal of developing models to predict receptor-mediated response [52,53].

2.2.1. (Q)SAR

Many attempts have been made in the past to relate, in a qualitative manner with SAR, or quantitatively with QSAR, molecular characteristics to some observed properties. Usually in the term (Q)SAR both approaches are included.

In the '60s a fundamental contribution was made by Corwin Hansch [54,55]. Hansch's paradigm was based on the study of congeneric series of chemicals. The activity, expressed in the logarithmic form, was assumed to depend on the substituent contributions to the parent compound in terms of hydrophobicity, electronic and steric terms. The biological relevance of these terms was correlated with the ability of a compound to penetrate into a biosystem and to reach the target site for the required interaction [56].

To move from the requirement of congeneric series, a large variety of molecular descriptors have been proposed over the years to encode the structural features of chemical compounds [57].

The workflow of the QSAR process can be schematically represented in Figure 2.1. The assumption behind the development of a QSAR model is that a quantitative relation exists between molecular features and the studied biological activity.

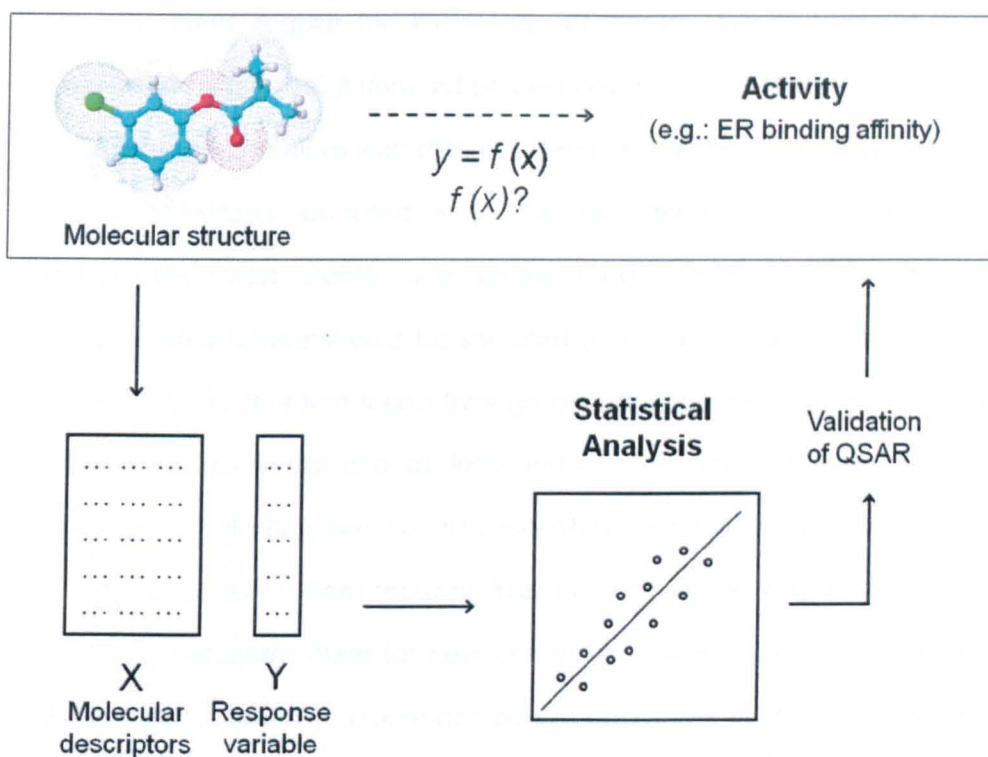


Figure 2.1 Schematic representation of the steps for developing (Q)SAR models.

To find this relationship the requested steps are:

- ① Calculation of chemical descriptors: As already mentioned, several types of descriptors can be used to encode different properties of chemicals (e.g.: electrostatic, hydrophobic, steric, topological, etc.) [57]. They can include experimentally calculated physico-chemical properties (e.g. boiling point), but in the majority of the studies they are constituted by a numerical evaluation of properties that can be computed on the basis of the chemical structure. A very practical approach for grouping the descriptor categories is to consider the structural information required to calculate them. Some kinds of descriptors encode very simple characteristics of the molecules that do not depend on the 3D conformation of the molecule and can be easily computed on the basis of 2D structure only. Other characteristics, such as

energetic terms, require the knowledge of the structure in terms of its 3D shape. This is not a straightforward process because each molecule can exist in multiple conformations with different levels of stability and occurrence. The solution commonly adopted is to use as reference conformation the energetically most stable one under fixed conditions. To obtain it, conformational space should be sampled (conformational search) and the global minimum structure found through geometry optimisation via common force fields, semi-empirical or *ab-initio* methods [58]. Then, 3D descriptors are computed on this structure. For 2D descriptors it is relatively easy to set up a procedure to make them reproducible, ensuring in this way that it will be possible to calculate them for new compounds later on, and to apply the QSAR model; however it is more complex to guarantee this for 3D descriptors. Firstly, the optimisation can include non-deterministic steps, and secondly it has been observed that sometimes small changes in the 3D conformation can have a large effect on the 3D descriptors [59]. Moreover what can be obtained is only a reference 3D structure, not necessarily representing the bioactive conformation assumed by a molecule in its interaction with the biological environment. Despite these limitations 3D descriptors have been used successfully in many studies [60,61] even though on large and heterogeneous datasets they have sometimes shown to be no better than 2D descriptors [62].

- ② Preparation of the Y-block variable: Along with the X matrix containing the independent variables, a Y matrix containing the target property or properties to be studied should be collected. If the target is a single activity this matrix consists of a single column containing the activity value associated to each chemical. The property to be modelled can be a continuous variable modelled in a quantitative manner or a categorical one modelled with classification techniques. One of the major problems in preparing the activity

data is that any algorithm adopted during the modelling relies on data to extract rules describing the activity trend. If these data are unreliable the resulting model will be misleading. So careful attention should be paid in preparing a dataset suitable for modelling purposes, by pruning ambiguous data. It should also be noticed that, despite all efforts in using only reliable data, there is an intrinsic variability within the experimental data that cannot be avoided, especially if they involve biological systems [59]. In this way a certain degree of noise is introduced in the system and the modelling step should distinguish between the relevant information contained in the data from the noise and redundancy introduced with either X or Y-block variables.

- ③ Statistical analysis: This is the central step of the modelling task. It includes pre-processing of data matrix, variable selection to include only statistically relevant descriptors, and the application of specific algorithms to find the relationship between chemical descriptors and the target property.

Pre-processing is a preliminary but essential stage where data matrix is pruned of redundant information, incomplete variables are deleted and the scaling procedure is applied to the dataset.

Selection of important variables can be done in two ways: hypothesis driven, including only variables considered a priori relevant for the endpoint to be modelled, or statistically driven using mathematical algorithms to search for the most relevant solutions. A stepwise approach or multivariate data exploratory methods such as Principal Component Analysis (PCA) may be used. However if the number of initial variables is too large these methods do not efficiently explore all possible combinations of variables and more sophisticated tools are required. Genetic Algorithms (GA), based on the Darwinian evolutionary theory, have been shown to be one of the most promising algorithms in this field. The best individuals in a population of models

are crossed over, merged, mutated and then iteratively evaluated against a fitness function which gives a statistical evaluation of the model performances. The variety of methods available to derive a model is also rather wide. A first distinction can be made between QSAR and SAR. While QSAR is searching for a quantitative relationship, SAR is a qualitative association between a molecular substructure and the presence or absence of a certain activity or the capacity to modulate that activity. The algorithms used for the modelling may vary depending on the kind of study: some studies deal with categorical target properties, which employ classification tools, and others with continuous variables, using regression approaches. Beside the more classical multivariate techniques that use linear methods, in recent decades neural networks (NN) have frequently been used as a non-linear statistical data modelling tool. NN is inspired by the way biological nervous systems, such as the brain, process information. It is a system of interconnecting neurons in a network, working together to produce an output function [63].

④ Validation of the model:

Whatever is the technique chosen to model a specific dataset, one of the most important issues in the QSAR field is that of validation of the models. When a qualitative or quantitative model has to be assessed, several characteristics of the model have to be analysed focusing in particular on three main aspects: (i) internal validation, (ii) prediction ability and (iii) applicability domain.

(i) Internal validation is based on the assessment of goodness-of-fit and robustness. The first concept applies to the ability of the model to describe the variation in the training set, while the latter provides an indication on the model stability in terms of how sensitive it is to perturbation in the training set. Commonly, for models based on continuous responses, the main statistical parameter to assess the goodness-of-fit is the coefficient of determination R^2 :

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}; \quad 0 \leq R^2 \leq 1$$

where y , \hat{y}_i and \bar{y} are respectively the observed, calculated and mean values of the Y dependent variable. The closer is R^2 to 1 the better the model is in fitting the data.

For classification models the quality of the model can be assessed by measuring its accuracy: the ratio between correctly classified compounds and the total number of compounds included in the dataset.

Robustness is usually assessed by the cross-validation procedure: the training set is iteratively perturbed in its composition by excluding one or more compounds and the other compounds are used to generate a model predicting the excluded chemicals with this sub-model. This procedure is usually repeated for all compounds. Statistical parameters similar to accuracy or R^2 (usually called Q^2 or R^2_{cv}) are then calculated based on the predicted values and should maintain a value comparable to R^2 . Randomisation of the Y-response can be also performed to evaluate whether with a dataset containing Y-scrambled responses the model statistics decrease significantly, as expected, or if not, it is an indication of chance correlation.

(ii) Traditionally QSAR models have been developed to describe a phenomenon, suitable to identify a rational relationship between a given parameter and the property. However, later on the emphasis has been put on the use of these relationships to predict the properties of unknown compounds and consequently different tools become necessary to avoid over-fitting. Indeed, in many cases, especially when many descriptors and complex algorithms are used, there is a risk of obtaining an over-fitted model that follows too close the behaviour of the training set not being able to generalize

and capturing the activity trend in a more general way. For these reasons statistical tools should prove the capability of the model to be valid in a general way, i.e. to be predictive upon compounds not used in development of the model. There is a debate in the scientific community on the most suitable way to assess the robustness and predictive performances of a model [64-66]. The use of an external set for the validation has been indicated in some cases as the most appropriate solution to assess the predictive power of a model [67] even though some difficulties may arise in designing it as representative of the training set, covering all its main structural and physico-chemical characteristics [68]. This issue is also related to the applicability domain problem.

(iii) The concept of applicability domain (AD) is an increasing consideration in the QSAR field due to the need to define better areas where is it possible to use the models practically with an increased confidence about the prediction so obtained [69]. It is based on the assessment of similarity for the new chemical to be predicted with the group of compounds used to develop the model. Among the available approaches there are series of chemometric tools based on the comparison of the descriptors used to develop the models of the new molecules to be tested, with those of the molecules in the training set. Another approach involves comparison of the structural features of the compounds in an *a priori* way, without necessarily using the descriptors selected in the models. In this case structures are encoded in fingerprints or by taking into account relevant fragments and using them to assess the similarity with the training set. A review of these approaches has been recently published [70].

Overall the validation should ensure that the model is statistically significant, reliable and robust against noise and data perturbation, and maintains its validity when the relationship is tested on compounds sharing similarity with

the training data, at least within a defined chemical space. For these reasons the judgment of model validity is based on the evaluation of a series of aspects, here summarised, sometimes assessed with a variety of different methods.

2.2.1.1. Models on ER with (Q)SAR

ER has been widely studied with QSAR techniques in respect of the EDs issue. Many of the studies were focused on the binding assay data. The datasets used in these studies are relatively heterogeneous in terms of both the number of compounds used - from a few dozen up to a few hundred - and the source for binding activity data - utilizing different species (rat, mouse, human, calf) and subtypes (alpha, beta or both mixed) -.

Some studies have employed structural features to discriminate binders versus non-binders in a SAR. The ability of compounds to bind the ER was associated by Fang *et al.* in a qualitative manner, to the presence of certain characteristics (e.g. a phenolic ring) [48] while Klopman *et al.* used the occurrence of certain groups among active or inactive chemicals to characterize recursively the most relevant fragments in the two groups in a semi-quantitative way [71].

Classification methods have also been used to categorize the data in two classes, employing different cut-offs to discriminate binders from non-binders or chemicals exhibiting marginal activity from strong binders. The most interesting approaches have employed Bayesian probability analysis [72], decision trees [73], soft independent modelling by class analogy (SIMCA) and spline-fitting with genetic algorithm [74].

Moving to quantitative studies, linear regression models have been produced in particular using PLS [61,75], a regression method based on the use of "latent variables" generated by a linear combination of the original set of descriptors. Other non-linear techniques have been explored and sometimes compared to

linear methods. One of them is a non linear QSAR technique which is based on the concept of molecular similarity and K-nearest neighbour principle [76,77]. Other non-linear methods have relied on the NN technique and different kinds of NN architectures have been explored, including probabilistic NN [78], back-propagation and counter-propagation NN [75].

To complete this overview, another approach has to be mentioned: it involves the use of multiple energetically reasonable conformations for each chemical to overcome the limitations of using a single 3D global minimum as reference structure to calculate descriptors [79]. Then, the distribution in the population of values for the descriptors calculated on the active compounds is compared with that obtained on the inactive compounds and it is used to derive a classification model.

Some studies have dealt with a different endpoint for oestrogenicity: instead of developing models focused on binding affinity data, the ER transactivation properties were investigated in terms of reporter gene assay or cell proliferation assay. Classification models have been produced by using classification trees [80], SAR approach [81] and a 3D-QSAR method based on Molecular Quantum Similarity Measures (MQSM) [82].

It is difficult to compare the results of different approaches since often these models relied on different datasets or at least on different validation parameters. Regarding the chemical information, it has to be highlighted that several of the paper here mentioned used 2D descriptors or even compared performances with 3D descriptors. The majority of these studies found that with 2D descriptors, which are simpler to be calculated and allow for a relatively faster analysis, it is possible to obtain comparable results with those involving more complex 3D descriptors [73,83,84]. This observation is somewhat surprising since the weakness of 2D descriptors is that they ignore the 3D nature of chemical interaction with a receptor, which is an important ingredient in the binding. A possible explanation is

that this kind of trend was especially noticed in more heterogeneous datasets or in classification studies where a lower level of detail in the chemical information is required to obtain for instance a binary classifier. This finding can surely be an advantage for having models more easily exploitable for new chemicals, and hence more useful in practical terms for prioritising new compounds. On the other hand, the interpretability of the selected descriptors is sometimes less explicit, especially if the models use a relatively large number of variables.

Comparing the different modelling approaches adopted, slight improvements were obtained with non-linear methods but often the statistical significance of the improvements was too limited compared to the increased complexity.

Some attempts have been made for improving the regulatory acceptance of these models, in particular in starting to address the problem of defining better the applicability domain [73,80]. A few of the models here briefly described will be analysed in more detail in the next chapter of this work, focusing especially on these aspects.

2.2.2. 3D-QSAR

3D-QSAR includes a variety of methods which basically differ from the classical QSAR analysis because of the descriptor types used within the modelling. These methods are based on the concept of Molecular Interaction Fields (MIF) [85]. Molecular features are derived by mapping the environment surrounding the molecules in terms of energetic interactions of various natures, mainly steric and electrostatic, but sometimes hydrophobic or hydrogen bonding potential may be included. This is obtained by placing the molecule within a lattice and calculating the interaction energies of this molecule with a probe (e.g. a sp^3 carbon atom) at each point of the grid, as shown in Figure 2.2.

Since field-based descriptors are directionally dependent, a very critical step in the 3D-QSAR analysis is the alignment in the Euclidean space of all the molecules

in the dataset accordingly to various methodologies of superposition. The alignment can be based on the electrostatic and/or steric field overlap, based on a common skeleton superposition, evaluating docking or crystallographic information, or based on pharmacophoric hypothesis.

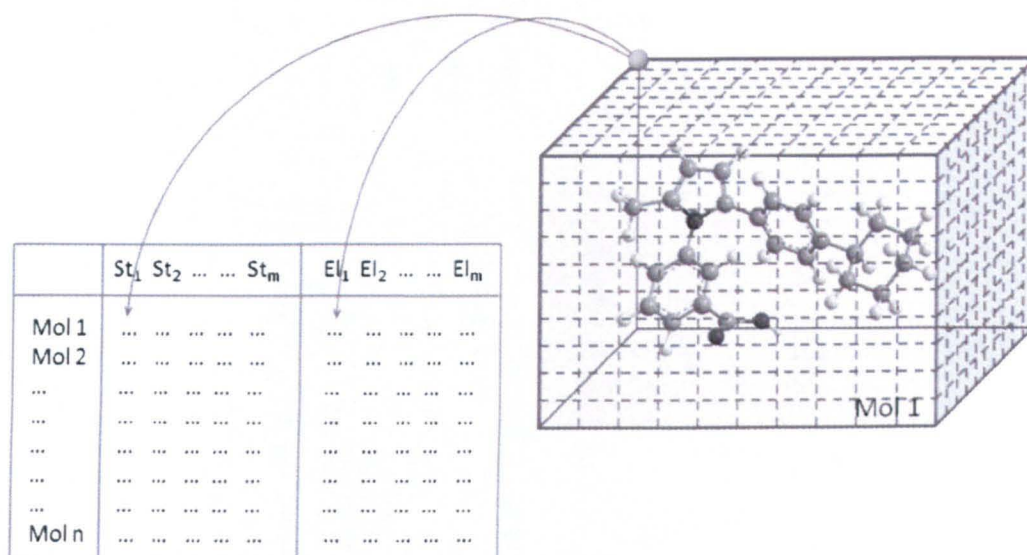


Figure 2.2 Representation of the lattice used to calculate field-based descriptors. 3D-QSAR descriptors are calculated by placing appropriately overlapped molecules within a lattice and calculating with a probe (placed in the upper left hand corner in the figure) the steric and electrostatic field contributions of the molecule at that point. Each matrix column contains the contribution to the field for each molecule at a specific grid point.

Each descriptor column is made up of the values of the interaction field assumed for the compounds included in the dataset in a certain point of the grid. Then these field energy terms are used as a very large pool of descriptors – hundreds or thousands – to search for a relationship with the property of interest, usually by a PLS analysis, thus the variation from molecule to molecule in the fields they generated in some areas of the grid are usually related to the variation in the modelled activity. Since these interactions are clearly placed in the 3D space surrounding the molecules, a regression map can be created by mapping back

into the box the regression coefficient of the model. Therefore the method allows the identification of the most important regions in the molecule responsible for modulating the target properties.

The most popular method in 3D-QSAR studies is CoMFA [86] but other methods have been developed based on different force fields adopted to calculate the energy terms, and considering more heterogeneous probe definitions to capture more complex interactions, for example CoMSIA [87] or GRID MIF [85], coupled with PLS analysis.

The most attractive feature of 3D-QSAR compared to classical QSAR is that the biological environment surrounding the molecules is taken into account even if only implicitly. This happens especially when direct interaction with a target macromolecule is considered (e.g. activity against a specific receptor) and a hypothesis about these interactions can be made on the basis of the derived relationship. The model can be interpreted so that the conformations of ligands are representative of the bioactive conformation assumed in the binding pocket of the receptor and the alignment represents the different possibilities of the molecule's binding to the receptor. For this reason the choice of the bioactive conformation and proper alignment are essential phases and often information on these characteristics is derived from crystallographic or docking studies.

Traditionally, linear PLS method has been used to derive CoMFA models since this method is tolerant to the inclusion of a large number of variables in the model, although other methods can be used to investigate nonlinear relationships, such as neural networks. However, including into the model a large number of variables, even if condensed in a few principal components, can increase the risk of chance correlation. For this reason the validation procedure is an essential step to test the model performances, particularly for the 3D-QSAR method. Commonly the cross-validation procedure is adopted and sometimes a randomisation test is included. Often the model performances are verified by evaluating the

prediction for new compounds, included in a test set. However neither results from cross-validation nor performances on the test set compounds can give a reliable estimation of model reliability if these parameters have been used to select the best architecture of the models. A more extensive discussion about these issues can be found in a review paper by Y. Martin [88].

The attention to the biological environment along with the ease of interpretability of chemical features through the regression maps are two of the major advantages of 3D-QSAR.

On the other hand, the principal drawback is the increasing complexity of the models, which requires 3D conformations, their alignment and a large number of variables. This can sometimes make it more difficult to reproduce a model, or at least to apply it to new compounds, if the alignment rules are too specific or are not suitable for other chemical classes, thereby limiting the range of chemicals that can be analysed with the model.

To overcome the limitations due to the superposition procedure some alignment independent extensions of 3D-QSAR descriptors have been developed that do not require aligned structure, such as VolSurf and GRIND derived from GRID/PLS [85].

2.2.2.1. Models on ER with 3D-QSAR

3D-QSAR methodologies have been widely applied in the study of receptor-ligand interactions, since the presence of a defined biological reaction site makes the determination of a proper alignment more likely.

Often modelling exercises have been conducted using both 3D-QSAR and classical QSAR approaches and comparing their outcomes based on the investigation of a common dataset [83,89]. For instance, a comparative study was performed using CoMFA and CoMSIA, classical 2D/3D descriptors, and fingerprint type descriptors that characterize chemical structures in a string of bits

indicating the presence or absence of specific 2D or 3D structural characteristics [83]. Another study proposed CoMFA in combination with 2D-QSAR methodologies based on fingerprints descriptors (HQSAR, and FRED/SKEYS), providing a helpful comparison of their predictive power [89].

Overall these studies demonstrated that for heterogeneous datasets 3D and classical QSAR approaches exhibit similar performances, not necessarily justifying the use of 3D-QSAR which implies an increased complexity. It has to be noted that this kind of assessment is often based on a limited set of validation parameters, provided in the original studies, and for this reason different conclusions may be derived if a larger pool of validation factors is used. On the other hand 3D-QSAR can provide a more easily interpretable model in terms of chemical features, especially if compared to other less intuitive and less transparent molecular descriptors.

3D-QSAR can be used to highlight differences in receptor affinity and to model the selectivity of the ligands for some receptor subtypes. Tong et al. [90] employed CoMFA maps to identify and differentiate the structural features of ligands responsible for the selective binding to ER alpha and beta. Although the receptor crystal structure is available, CoMFA provides an additional source of information about the receptor from the perspective of the ligands.

Other successful applications were for series of relatively homogeneous compounds where proper alignment rules can be more easily detected. This approach is commonly used also within the pharmaceutical industry for optimising the characteristics of a lead compound among homogeneous series. CoMFA, CoMSIA and HQSAR were used to investigate a series of Bisphenol A analogues, considering not only the binding but modelling also other properties such as transactivation potency [91]. Again, statistical performances of the different methods seem similar even if their respective outcomes are more complementary than alternative. Interestingly this study provided an example

where instead of the minimum conformation, the most probable bioactive conformer identified by other *in silico* simulation was used, through virtual docking.

2.2.3. VIRTUAL DOCKING

Virtual docking is a methodology to predict computationally the binding between two molecules, usually a protein and another macromolecule (protein or DNA) or a small molecule (ligand). Here the focus will concentrate on protein-ligand docking as a tool to estimate the interaction of chemical compounds with biological target sites.

To perform this kind of study the chemical composition and 3D spatial organisation of the protein should be known along with the identification of the cavity defining the binding site of the protein, whose position and shape are used in the docking process. Usually the most suitable source is the structure solved through X-ray crystallography. If this is not available, a structure determined with NMR spectroscopy or by homology modelling may be used. The latter method consists in the reconstruction of the 3D shape of the protein of interest from other proteins, whose structure is known, that share similarity in the aminoacid sequences. The protein structure so obtained can be less reliable compared to the other methods mentioned above.

It should be noted that using as starting point the crystal structure of a ligand-receptor complex, means that the procedure begins from a single, low-energy, snapshot of an actual dynamic biological system. Then the first task to accomplish is to sample the conformational space of possible energetically reasonable structures of the protein-ligand complexes. This is not a trivial process since the conformational space to be examined is huge due to the intrinsic flexibility of both the ligands and the protein, including further forms of plasticity that can be introduced upon the mutual recognition between protein and ligand

by the induced fit process [92]. Nowadays the majority of the docking programs take into account ligand flexibility in contrast to rigid docking (where both the ligand and the protein are considered as rigid bodies), but protein flexibility is not yet fully integrated into docking protocols and is taken into consideration only marginally. Among the methods employed to perform the searching strategy there are molecular dynamics simulations, Monte Carlo methods, genetic algorithms and fragment-based methods [93].

Once a pool of ligand-protein complexes has been generated, scoring functions are used by docking programs to indicate the likelihood that the orientation possesses a favourable binding interaction. The scoring function provides an estimation of the Gibbs free energy of binding, released when ligand and receptor bind, to evaluate their likely stability as a complex (see Figure 2.3).

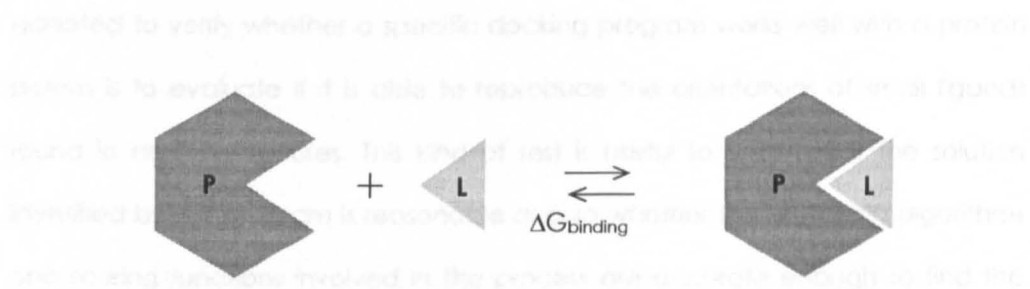


Figure 2.3 The process of mutual recognition between a protein (P) and a ligand (L) is governed by the Gibbs free energy of binding, released when ligand and receptor bind

For each orientation $\Delta G_{\text{binding}}$ can be used to assess whether a favourable opportunity of binding may exist through the following equation:

$$\Delta G_{\text{binding}} = RT \ln K_D$$

where R is the gas constant, T the absolute temperature and K_D the dissociation constant.

The dissociation constant can then be directly linked with the inhibition constant (K_i) or the inhibition concentration (IC_{50}) obtained in an *in vitro* binding assay.

Consequently, a proper scoring for the docked orientations is the second crucial step in the docking process. The scoring functions may rely on force fields to calculate energies or on knowledge-based or empirical functions (including QSAR relationships) [94].

Although fairly accurate ways for estimating these energies exist, based on free energy perturbation or thermodynamic integration methods, this level of accuracy in practice can be used only in a few, very focused studies. On the other hand, the most attractive application of virtual docking is its use in virtual high throughput screening (HTS), whereby large virtual libraries of compounds are reduced in size to a subset, which, if successful, includes molecules with high binding affinities to a target receptor [95]. This approach is often used in the drug discovery process to identify new lead compounds. One of the tests commonly adopted to verify whether a specific docking program works well with a protein system is to evaluate if it is able to reproduce the orientations of small ligands found in crystal structures. This kind of test is useful to evaluate if the solution identified by the program is reasonable and so whether the searching algorithms and scoring functions involved in the process are accurate enough to find the right solution (an orientation close to the experimentally determined one) among many others and to score it properly, recognizing it as one of the favoured orientations. This preliminary evaluation does not offer enough guarantee about the final docking output precision. For applications in virtual HTS the scoring function used to approximate the energies should be fast enough to accomplish this task in a reasonable time, but this is associated with a low level of accuracy. The typical level of accuracy reached by docking programs in virtual HTS does not allow a direct correlation of their scores with the binding affinity. Moreover they are hardly able qualitatively to rank the compound order properly in respect of binding strength [96].

Normally the outcomes obtained with this task can be measured with the enrichment factor so that the subset selected with the docking procedure contains a larger number of compounds showing affinity for the studied receptor. A practical comparison of performances and features of the most popular docking programs can be found in the literature [97].

The calculation complexity of virtual docking, which also encodes the protein structure, is clearly larger compared to the use of ligands alone, due to the larger number of atoms involved in the calculation and the need for proper force field parameterisation for the interaction of small molecules with aminoacid residues. However, once the experimental protocol is set, docking methods are fast enough to screen very large libraries of chemical compounds in a reasonable time.

Some limitations may arise from inaccurate energy estimation. Docking still remains effective in drug design, since it improves performances compared to random selection of possible hits, as indicated by the enrichment factor, whilst its application as a pre-screening tool targeted at the hazard assessment of less enhanced binding activities has been tested less often.

This latter application may be more critical since in this case there is less tolerance of false negatives while virtual HTS is commonly biased by the intensive use in drug design where the attention is more focused to control the number of false positives. Some interesting applications of the EDs issue can be found in literature and are analysed below.

2.2.3.1. Models on ER with docking

Some docking studies have been published addressing more or less specifically the issue of EDs. In some of these investigations a quantitative assessment of the binding was provided with the docking approach for a limited number of compounds under investigation.

Some encouraging outcomes of docking simulations have been obtained with molecular dynamics and linear interaction energy methods to measure the interaction energies fairly accurately and link them with the experimental binding. The validity of the approach has been also tested for predicting a few new compounds [98].

In another study a small group of steroids, phytoestrogens and PCBs were docked into the alpha ER and into a homology model of ER beta receptors. The calculated energies correlated very well for the investigated chemicals with the experimental binding affinities for both subtypes, recognizing different selectivity for the two subtypes [99].

For a larger set of heterogeneous compounds, binding affinities were correlated with docking scores less accurately but the approach can be considered successful for evaluating the enrichment factors in the screening. Interestingly, to account for receptor flexibility a subset of receptor crystal structures were used in parallel in the docking process and this approach increased the accuracy gained in comparison with the use of a single complex [100].

Other studies demonstrated that the docking methodology failed to provide enough accuracy in estimating binding strength as previously explained, but confirmed its validity as a complementary tool to develop more powerful 3D-QSAR models. Thus, the use of docking conformers to generate a more biologically plausible alignment creates a link between the two methods [101].

Some authors showed that, although docking itself failed to predict the binding affinities quantitatively, the use of docked orientations for the ligand and, with a minor influence, the inclusion of scores as additional parameters, allowed the development of a superior model compared to the classical CoMFA model using lowest energy conformer for ER alpha. However, this approach failed for ER beta [102].

There are also examples where direct calculation of binding energies based on docking runs gave poor results compared to classical CoMFA models, but the use of docked orientations instead of minimum energy conformer significantly increased the performances of 3D-QSAR, including encouraging activity prediction of new chemicals [103].

The docking approach has been used as a starting point to develop multi-dimensional QSAR model for ER [104]. The multiple dimensionalities were given by the inclusion of multiple conformations (4D), induced fit (5D) and solvation effects (6D). Receptor surrogates were derived by mapping the different properties on a surface surrounding the molecules and selecting the most appropriate ones to assess the binding affinity using GA. Very good quantitative results were obtained with this method for a large dataset of a hundred heterogeneous compounds.

2.3. CONCLUSIONS

Different ways for assessing the properties of chemical compounds *in silico* have been illustrated here. The characteristics of the different approaches are summarised in Table 2.1. The first distinction can be made between receptor-dependent and independent methods.

Traditionally, in environmental and health safety the phenomena under evaluation are more general ones and do not necessarily represent the explicit interaction with a well characterised, specific receptor (e.g. systemic toxicity or carcinogenic process); thus receptor-based methods can be applied only in some circumstances such as for investigating NR interactions or metabolic processes mediated by the cytochrome P450 family. In other contexts, when more general multi-step toxic effects of chemical compounds are studied, or a defined mode of action cannot be recognised, the modelling must rely on the chemical structure alone. In these cases QSAR and 3D-QSAR can be applied also

when the target macromolecule is not known or toxic effects cannot be linked with a specific receptor.

Table 2.1 Overview of the methods that can be adopted to study the biological effect of chemical compounds.

	TECHNIQUES		
	Receptor-independent		Receptor-dependent
	QSAR	3D-QSAR	Docking
DESCRIPTION	(Q)SAR searches for a qualitative or quantitative relationship describing the influence of small molecule features, encoded in molecular descriptors, in producing a certain biological effect through a statistically significant model.	It uses field-based descriptors to represent the energetic environment of the molecules. It requires 3D conformers which should be properly aligned.	The binding affinity of a ligand with a receptor is inferred by an energetic evaluation of the complex through scoring functions.
PROS	Quite fast and reproducible especially if descriptors depending only on 2D characteristics of the molecules are used. Widely applicable to new compounds: although the model itself can be complex to be generated, easy rules can be derived.	Considers the biological environment surrounding small molecule although implicitly. It allows for a visualisation of most important molecular characteristics through the regression maps.	Closer to reality: It explicitly includes a description of the biological macromolecules responsible for the observed activity. It allows a deeper mechanistic understanding.
CONS	Less realistic: It ignores the 3D biologically active conformation of the ligand and ignores the chemical structure of biological macromolecules responsible for producing the effects.	Requires 3D conformers and it is generally very sensitive to the alignment procedure.	It allows modelling the binding strength of ligands with proteins but not more complex and global biological effects. The accuracy that can be gained often permits only qualitative output or ranking of compounds for their activity.

Moving from classical QSAR to 3D-QSAR to docking there is an increasing attention to the biological side and this increases the "biological plausibility" of the results obtained by statistical methods.

Explicitly introducing a description of the biological receptor increases the accuracy of the biological environment description, and can provide more insights of the mechanistic interpretation. On the other hand, the increased complexity (e.g. the computational one), although it may improve the comprehension of the studied process, often does not produce more statistically significant models, indicating that the noise introduced in the system may be greater than the additional informative content.

It is not worth estimating *a priori* whether one technique is superior to another. What is frequently observed is that depending on the problem to be addressed some techniques can be more beneficial than others. Moreover it has been proven that complementary outcomes can be derived, so that a more complete description of a phenomenon can be obtained by integrating different techniques.

Literature sources describing relevant modelling exercises for endocrine disrupters, in terms of oestrogen receptor interferences, have also been reviewed, focusing the attention on possible applications to screen large chemical inventories for environmental and health safety.

In the remaining part of this research project, since the main goal was to provide some widely applicable models, the focus will be on classical QSAR tools.

PART II

VALIDATION OF EXISTING MODELS

CHAPTER 3

CRITICAL ASSESSMENT AND SELECTION OF EXISTING MODELS FOR EDs ACCORDINGLY TO THE OECD PRINCIPLES

Based on the literature review conducted at the beginning of this research, a number of interesting models were identified. Therefore the first part of this project was intended to analyse the existing information and models prior to developing new ones in order to: i) assess what level of accuracy was already obtained and ii) identify the most promising methods to be further explored.

Even though a huge number of models for oestrogen receptor binding have been published, developed using many different algorithms, little attention has generally been paid to the practical use of these models. More efforts are now required not only to derive new models but also to have a deeper understanding of their applicability to real world problems and particularly in relation to the regulatory use of QSAR [45,105-107]. In order to have a model accepted and used in the context of a regulatory framework five principles have been identified by OECD² as important steps for its validation and acceptance; in particular the following pieces of information to ensure the reproducibility of the model and its robustness and predictivity have to be provided:

- 1) a defined endpoint,
- 2) an unambiguous algorithm,
- 3) a defined domain of applicability,
- 4) appropriate measures of goodness-of-fit, robustness and predictivity,

² http://www.oecd.org/document/23/0,2340,en_2649_34365_33957015_1_1_1_1,00.html

- 5) a mechanistic interpretation, if possible.

Along with the precise definition of the protocol, ECVAM suggests that other steps are required to ensure QSAR validity from a regulatory point of view [108]. Basically it should be verified that the model itself is transferable by other scientists who can reproduce the model, evaluate its variability and apply it to new compounds.

This kind of external assessment may be very difficult to perform since it should rely on information made available in the public domain. On the other hand this practical proof of concepts is the only way to evaluate to what extent a given model is really applicable and suitable to address the specific activities taken into consideration.

The aim of the specific task presented here, was to identify relevant criteria for selecting promising (Q)SARs to perform an independent evaluation for a subset of interesting models. Due to the intrinsic difficulties and constraints mentioned earlier, it would not be possible to evaluate all the models identified in literature so this assessment intends to analyse essential information concerning the (Q)SAR models under investigation, and to select a reasonable subset of models to perform this external judgment.

3.1. CRITERIA FOR MODEL SELECTION

Two parallel evaluations were conducted to identify relevant criteria for model selection; first of all, from a theoretical point of view, the OECD principles were analysed in respect of the current status of researches on endocrine disrupters. In a second instance the reverse process was followed, considering the concrete information available in the models identified in the literature and evaluating how they fit within the OECD principles and the areas needing a further independent assessment.

3.1.1. PRELIMINARY EVALUATIONS OF THE OECD PRINCIPLES

Firstly some general considerations about the OECD principles were derived, evaluating their translation from a theoretical point of view to their practical application by considering the overall pool of models identified in the literature.

1st OECD principle: a defined endpoint. Regarding this principle it is interesting to note that there is great interest in endocrine disruptors from a regulatory point of view but validated tests and guidelines do not already exist.

Some preferred protocols for testing have been suggested by ICCVAM [24] but overall the available models in literature rely on existing, available data. For instance, most workers chose to investigate endpoints related to ER compared to other NRs; this is probably due to the interest on ER as a pharmaceutical target, and because of an earlier development and standardisation of the test procedures for ER, that made available a larger number of data to model.

So, the scientific relevance of this endpoint is certainly clearly characterised but it can be more difficult properly to identify its relevance in the context of the regulatory framework.

For instance, the receptor binding affinity (the endpoint most often studied), represents just the first step in a cascade of events and other endpoints downstream of receptor-ligand interactions may be more suitable properly to evaluate the effects. Moving from the binding experiments to the transcriptional activation properties or even to the *in vivo* characterisation of the effect, the biological relevance of the endpoint surely increases. However some criticisms may arise in defining whether these endpoints are appropriate to assess the risk posed by EDs, since it is extremely difficult to evaluate the significance of low-dose effects of weak xenoestrogens [109].

All these considerations are behind the development and validation of QSAR models but should be taken into account in the future when more details about the toxicological perspective become available.

2nd OECD principle: an unambiguous algorithm. This principle refers to all considerations about the equation encoding the final model itself but it includes also a definition of other characteristics such as the descriptors used in the model. Some critical aspects for the application of this principle can arise from both the chemical and mathematical side. From the chemical point of view it is crucial that the method to calculate the descriptors is disclosed so that it can be unambiguously applied to new chemical structures. This can be especially difficult for 3D-QSAR models, strongly depending upon the alignment, since it should be verified that the alignment rules are reproducible and applicable to new chemicals. Regarding the explicit definition of the equation or rules describing the final model, they can be easily accessible if the model itself has a simple structure, or they can be more complex and so more difficult to reproduce.

3rd OECD principle: a defined domain of applicability. The third principle, dealing with the definition of the applicability domain, has been explored in literature only marginally and no precise definition of the AD is available for almost all the models in literature. A related issue that on the other hand has been sometimes addressed is the assessment of the portion of the chemical space of interest that can be evaluated with the proposed model, which reflects the practical value of the model as a screening tool.

4th OECD principle: appropriate measures of goodness-of-fit, robustness and predictivity. Within this statement all issues concerning the model statistics

themselves can be found, including the access to background information regarding training and test set, modelling and validation procedures. Even though within this principle a number of essential statistical measures for a realistic assessment of the model are included, in practice all models published in literature deal with these topics in a very heterogeneous way and adopt *ad hoc* parameters to assess their performances, making it difficult to compare them with the others.

5th OECD principle: a mechanistic interpretation, if possible. In the specific case of oestrogen receptor interaction, a number of pieces of evidence are coming from other fields of research that can support a mechanistic interpretation of the models sometimes prior to model development, guiding the descriptor selection, more often *a posteriori*, after the descriptors have been selected to rationalize their importance and justify their use.

In conclusion, there are certainly some critical areas that require a deeper evaluation of the OECD principles with their application to some concrete examples.

3.1.2. DEFINITION OF THE CRITERIA SUITABLE FOR SHORT-LISTING OF QSAR

To highlight critical factors that can prevent the specific models to be independently assessed and validated, the OECD principles were re-evaluated before analysis from a more practical point of view.

The main criteria used to select the (Q)SAR models to be validated were:

- a. Experimental biological data
- b. Chemical information on structures
- c. Chemical information on descriptors
- d. Chemical domain

e. Modelling approach and feasibility of the model

In particular the rationale behind the choice of these criteria was the following:

- a. Experimental biological data should be reported for each chemical in the exact terms of the modelled output or detailing the necessary transformation (e.g.: range of activity for defining classes, logarithmic transformation of the continuous values). Data can be reported from a different source from that used to develop the model, but should be available in the public domain. This will be a minimum requirement to consider the model for the next steps.
- b. Chemical information on the structures should be given; at least in terms of sketched 2D structure or chemical name to identify the substances under evaluation in order to be able to derive the data matrix associating chemicals to the activity data. Again, this will be a minimum requirement to consider the model for the next steps.
- c. Chemical information on the descriptors and how they were calculated should be given. If the calculated descriptors and in particular the descriptors used to build the model are only generally defined (e.g.: reactivity parameters without any further detail) they cannot be recalculated so the model itself is not reproducible. A discrete scale to assess this parameter is necessary, in particular some primary characteristics have to be considered, for instance: i) the ease of reproducing the descriptors (e.g.: CoMFA descriptors depending on the alignment) and ii) if the table with the calculated descriptors is given, in order to check for consistency with the new recalculated values.
- d. A specific parameter to consider is the chemical domain, identifying with this term a different concept from the domain of applicability. Even in case of formal acceptability, i.e. the model is sufficiently transparent, its utility can be limited because it is too restrictive in its chemical space. In this way it is possible to encode for the real utility of the model for screening industrial

chemicals, based on a qualitative assessment of the chemical diversity and the representation of chemical classes relevant from an environmental and toxicological point of view. This kind of approach is essential when dealing with ER models to evaluate their utility in terms of endocrine disrupters screening since many models were more pharmaceutical-oriented: developed using small datasets of very similar synthetic candidate drugs, whose utility for a large screening of industrial chemicals will be dubious.

- e. This issue refers to several aspects connected with the modelling task. Basically it encodes for the availability of sufficient information on the parameters of the model, which allow for it to be reproduced. A further point relates to the feasibility of the modelling approach; it means that easier models are more acceptable, but more complex models can be selected as well. Finally an overall evaluation of the models is given in terms of their performances, simply to avoid the selection of models which have already been shown to be relatively weak in their performances or have failed some validation attempts in the original work.

Information on the mode of action and on the applicability domain in a proper sense was not considered for the reasons explained in section 3.1.1.

3.2. SELECTED MODELS

Considering all criteria here exposed three models were selected from the literature [48,61,73] as candidates for an independent and complete validation (see Chapter 4). The selected models are reported in Table 3.1 with a brief description of their characteristics.

The intrinsic characteristics of the criteria that guided the model selection ensure that models using relatively large and diverse dataset were preferred, and this will help in better evaluating the applicability domain definition. Enough supporting

information is available for these models in the original publications or are easily accessible (i.e. biological data and descriptors originally used). Moreover it maximised the selection of models employing different techniques, e.g. regression and classification, quantitative or qualitative SAR to take into account also this variable.

Table 3.1 Models selected for external validation.

	Studied endpoint	Modelling technique
Model 1 Ref. [61]	ER Relative Binding Affinity (RBA) for rat Prediction of continuous values for active compounds	MLR equation using 8 descriptors (6 2D descr. + 2 3D descr.) $n = 131$
Model 2 Ref. [48]	ER classification model for rat RBA Prediction of two classes: active (any detectable activity) versus inactive	2D descriptors (Molconn-Z) used to develop a Decision Forest model (combination of multiple Decision Trees) $n = 232$: 131 Active and 101 Inactive
Model 3 Ref. [73]	SAR model for rat RBA Prediction of two classes: active (any detectable activity) versus inactive	This SAR model identifies six features used in a flowchart to discriminate active versus inactive compounds $n = 232$: (131 Active and 101 Inactive)

CHAPTER 4

INDEPENDENT VALIDATION OF SELECTED (Q)SARS ON THE BASIS OF THE OECD PRINCIPLES

Several aspects of the three models selected in the previous Chapter have been considered in order to perform an independent validation of their performances accordingly to the OECD principles.

In particular the following characteristics were considered in detail:

- 1) The transparency and reproducibility of all steps were carefully evaluated;
- 2) Performances in prediction with new compounds identified in literature;
- 3) Assessment of the AD;

These issues do not completely cover all the OECD principles for QSAR validation but consider them from a more practical point of view analysing those aspects that can limit the application of the models. In this way it will be possible to assess how easily and quickly the models can be transferred from the original developers to other users and their predictive performances.

4.1. MATERIALS AND METHODS

4.1.1. *SELECTED MODELS TO VALIDATE*

4.1.1.1. Sources for experimental activity data of the models

Experimental activity data, used in all three original models in Table 3.1, are from the NCTRER dataset (FDA's National Center for Toxicological Research - Estrogen Receptor Binding Database). It contains oestrogen receptor binding affinity data relative to 17 β -estradiol (RBA), calculated as the ratio of IC₅₀ values of 17 β -

estradiol and the test substance, multiplied by 100. A group of 232 chemicals was tested in a radioligand competitor *in vitro* assay, using rat uterine cytosol. It consisted of 131 active compounds (activity ranges from – 4.5 to 2.6, log unit) and 101 inactive compounds with no detectable activity in the assay. Toxicity data were published in several papers and publicly accessible through internet sources¹ [110,111].

Model 1 used only the subset of 131 active compounds while the others provide a binary classification for the entire dataset (131 active vs. 101 inactive chemicals).

4.1.1.2. Descriptors

The chemical information was treated differently in the three models. *Model 1* used 8 molecular descriptors calculated with TSAR [112] and QSARis (now MDL QSAR) [113], two of them being quantum-chemical descriptors depending on the optimisation of the 3D structure.

About the description of molecules, the SAR model (*Model 2*) identifies the following six features used in a flowchart to discriminate active versus inactive compounds:

- F1 Ring: Presence or absence of a ring in the chemical structure;
- F2 Aromatic Ring: Presence or absence of an aromatic ring in the chemical structure;
- F3 Phenolic Ring: Presence or absence of a phenolic ring in the chemical structure;
- F4 Heteroatom: Presence or absence of a H-bond capable heteroatom (O,S,N) attached to a non-aromatic ring structure;
- F5 Phenol 3n-Phenyl: Presence or absence of a phenolic ring linked by 1-3 bridging atoms (C or O) to another aromatic ring system;

¹ <http://edkb.fda.gov/webstart/edkb/index.html>

- F6 Other Key Features: Presence or absence of a key structural feature conferring activity, more precisely: H-bonding ability, Precise O-O distance (11 Å), Rigid structure, Steric moieties mimicking 7 α and 11 β position of E2 and Satisfactory hydrophobicity (LogP).

In Model 3 only 2D descriptors were used for the dataset named ER232, and these were computed using Molconn-Z¹, version 4.07. After removing descriptors that were constant across all chemicals in a data set, more than 270 descriptors remained and were used in model development.

The spreadsheets with the originally calculated molecular descriptors or features for all three models were made available directly by the authors, or were freely accessible through the internet.

4.1.1.3. Mathematical formulation of the models

Model 1

In the paper by Ghafourian & Cronin [61] several linear models, based on MLR and PLS techniques, are reported. Among them, the stepwise MLR model was at the same time the least complex (fewest descriptors) and most predictive one and was chosen for this validation exercise. Confidence limits are also reported.

$$\begin{aligned} \text{Log RBA} = & +0.14(\pm 2.6) + 0.593(\pm 0.05)N_C + 1.94(\pm 0.28)I_{\text{phenol}} - 0.0013(\pm 0.0002)W - \\ & + 14.9(\pm 2.1)^6\chi_{\text{ch}} + 0.985(\pm 0.26)E_{\text{HOMO}} + 0.0743(\pm 0.02)E_{\text{torsion}} - \\ & + 0.262(\pm 0.08)^3\kappa_a + 0.414(\pm 0.08)N_{\text{halogen}} \end{aligned}$$

$$n = 131 \quad s = 0.965 \quad R^2 = 0.723 \quad R^2_{\text{cv}} = 0.679 \quad F = 39.8 \quad p = 0.000$$

For a complete definition of descriptors see Table 4.1.

¹ <http://www.edusoft-lc.com/molconn/>

Model 2

The SAR analysis is shown in Figure 4.1, where the six key features of Model 2 are inserted in a flowchart, whose flux identifies compounds likely or unlikely to bind to the ER.

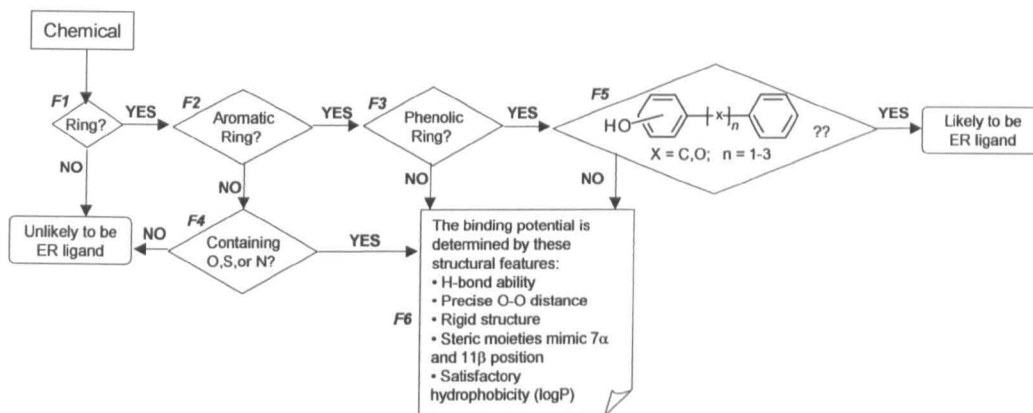


Figure 4.1 The flowchart shows the rules for the SAR model as reported in the DSSTox database [111].

Model 3

The Decision Forest (DF) method was used to develop classification models. It is a consensus modelling technique that combines multiple Decision Tree models, maximizing the diversity of the descriptors included in each tree. The final model published on the ER232 dataset combines six trees and is based on about 80 descriptors.

4.1.2. INTERNAL VALIDATION AND MODEL REPRODUCIBILITY

Following the OECD principles, models for regulatory purposes have to be transparent. To accomplish this requirement an important step is that of reproducing and verifying the models and the associated statistical parameters. In this context the recalculation of chemical descriptors is very important to verify the model reproducibility and the likelihood of obtaining the descriptors using different tools employing slightly different procedure/software (sensitivity study).

4.1.2.1. Re-calculation of the models and statistical assessment

The equation of *Model 1* was recalculated with Statistica [114] starting from the original descriptor spreadsheet; rules for *Model 2* were implemented directly in MS Excel while to re-calculate *Model 3* the original program developed by the authors was used, as it is freely downloadable from the web¹.

For the quantitative model, R^2 and R^2_{cv} parameters were used to assess their statistical validity and compared with the parameters provided by the original publication.

Several parameters, including Cooper statistics [115], have been used for evaluating performances of classifiers based on the number of correctly classified active (true positive = TP) and inactive (true negative = TN) compounds and the number of misclassified active (false negative = FN) and inactive (false positive = FP) compounds:

$$1. \text{ Prediction Accuracy} = \frac{(TP + TN)}{Tot} \times 100$$

$$2. \text{ Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$3. \text{ Specificity} = \frac{TN}{TN + FP} \times 100$$

$$4. \text{ Matthews correlation coefficient (MCC)} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$$

MCC [116] is a robust measure to evaluate a method that accounts for unbalancing (both over-prediction and under-prediction). It is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. MCC is a number between -1 and 1. A coefficient of 1 represents a perfect prediction, 0 an average random prediction and -1 the worst possible prediction. Thus, higher the correlation coefficient the better is the prediction performance.

¹ <http://www.fda.gov/nctr/science/centers/toxicoinformatics/DecisionForest/index.htm>

$$5. \text{ Performance with respect to random prediction } (S) = \frac{(TP + TN) - R_{Tot}}{Tot - R_{Tot}}$$

$$\text{where: } R_{Tot} = \frac{(TP + FP) \cdot (TP + FN) + (TN + FP) \cdot (TN + FN)}{Tot}$$

With this parameter the accuracy is compared with respect to a randomly generated prediction (R_{Tot}) and to the normalised percentage better-than-normal (S) [117].

4.1.2.2. Sensitivity study

For *Model 1*, which relies on a limited number of descriptors, some of them depending on the 3D optimisation process, a more detailed sensitivity study was performed to assess descriptor variability and their influence on the model performances.

Chemical descriptors were re-calculated starting from chemical structures publicly available through the Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network [111].

Compared to the original procedure and software other slightly different methods were used, as described and compared in Table 4.1.

This task permitted the evaluation of how much influence the procedural tasks have in introducing a certain amount of variability in the model and allowed the testing of whether the procedure used for calculating chemical descriptors for the new compounds in the external set can have a large influence on the predicted activities.

Table 4.1 List of descriptors used in *Model 1* and explanation of the procedure used to recalculate them.

Name	Description	Type	Original software	Recalculation procedure
N _C	Number of carbon atoms	Constitutional (1D)	Tsar	Calculated with CODESSA [118] from the sdf file provided in the DSSTox database
N _{halogens}	Number of halogen atoms	Constitutional (1D)	Tsar	Calculated with CODESSA [118] from the sdf file provided in the DSSTox database
I _{phenol}	Indicator variable for phenol group (presence/absence)	Functional groups (1D)	Tsar	Visual inspection
W	Wiener Topological index	Topological (2D)	Tsar	Calculated with CODESSA [118] from the sdf file provided in the DSSTox database
³ K _α	3rd order kappa alpha shape index	Topological (2D)	QSARis	Calculated with QSARis [113]
⁶ χ _{ch}	6th order simple chain molecular connectivities	Topological (2D)	QSARis	Calculated with QSARis [113]
E _{torsion}	Energy of torsion for the molecule by COSMIC Force Field	Quantum-chemical (3D)	Tsar	CORINA conversion in 3D structure, full COSMIC optimisation with TSAR [112]
E _{HOMO}	Energy of the highest occupied molecular orbital calculated by VAMP (using the AM1 Hamiltonian) in TSAR	Quantum-chemical (3D)	Tsar	Calculated with two methods: or 1) CORINA conversion in 3D structure, full COSMIC optimisation and after that AM1 optimisation performed with TSAR (HomoTSAR) [112] or 2) automatic optimisation with MOPAC AM1 in the OpenMolGRID environment [119] and extraction of the quantum-chemical parameters with Codessa (HomoCODESSA) [118]

4.1.3. EXTERNAL VALIDATION PROCEDURES

Model 1

The literature was investigated to identify other suitable sources of ER binding activity data in order to perform an external validation with new chemical compounds. This task implied an assessment of the correlation existing between new data and those used to develop the model for compounds already present in the NCTRER dataset.

For *Model 1* three further groups of compounds were used as external sets:

- A first check was done with 95 inactive compounds belonging to the NCTRER dataset. Of course this test set is only partially representative (from the activity point of view) of the original training set, because it involves only inactive compounds. On the other hand, this set was initially chosen because it was obtained under the same experimental conditions as the compounds used in the training set, and for the availability of the originally calculated descriptors. The model equation was implemented in MS Excel.
- A further 45 compounds were extracted from the EDKB database¹ and their activity data were taken from an aggregation of data from the literature for ER RBA [73]. No further details about test conditions were provided. Chemical descriptors were calculated on the basis of the procedure described in paragraph 4.1.2 and the model equation was implemented in MS Excel.
- In the literature another possible source of data was made available by the Japanese Ministry of Economy, Trade and Industry (METI) [120]. Some limitations are implicit in the choice of these data: they refer to human ER, not to rat, and data were obtained using a specific subtype of the receptor (ER α) while the data source for *Model 1* is the NCTRER dataset which used rat uterine cytosol ER. In uterine cytosol α subtype is predominant but ER β is also present and this can affect results, especially for compounds having selectivity for one of the subtypes such as phytoestrogens.

For all these reasons, before these data could be used as a test set a mathematical evaluation of the correlation between the two sources of activity data was conducted. Results are shown in Figure 4.2: the first series of data contains only the 68 compounds with defined activity values for both databases (indicated as defined series). The second series contains compounds inactive in

¹ <http://edkb.fda.gov/webstart/edkb/index.html>

at least one of the two sources. The correlation line reported in the graph is calculated on the basis of the "defined" series alone.

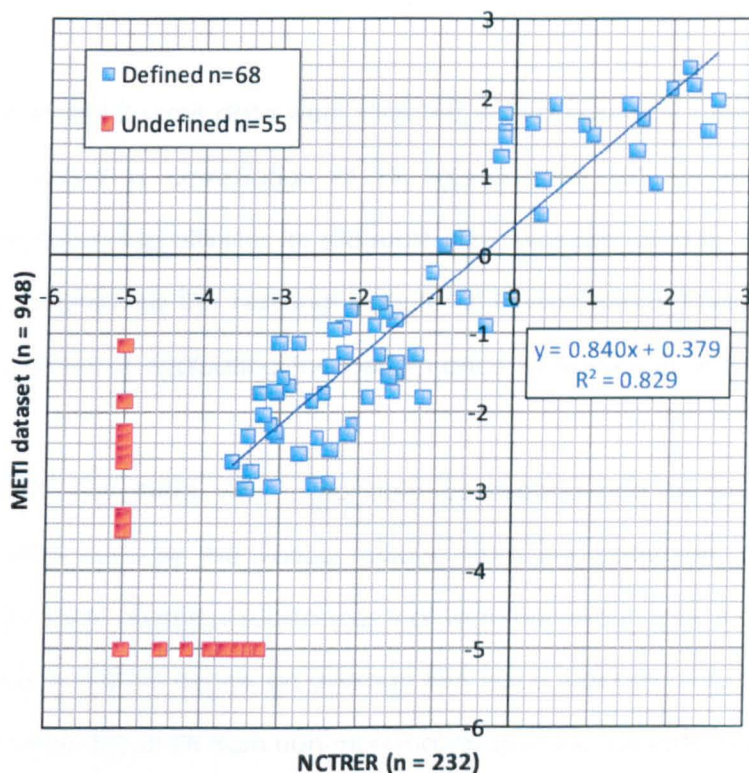


Figure 4.2 Correlation of the activity data found in the NCTRER database, used to develop *Model 1*, and the METI database. In the graph – 5 value was arbitrarily assigned to inactive compounds. The "undefined" series contains 9 compounds which are inactive in the Japanese database but active in the NCTRER database (with activity values < -3.25), 36 compounds inactive in both databases and 10 compounds inactive in the NCTRER database but active in the Japanese database (with activity values < -1.15).

Overall there is a relatively good correlation among the two database but values are somewhat scattered around the ideal correlation line (within ± 1 log unit). A lower correlation is observed if inactive compounds are considered as well in the analysis. Bearing in mind these limitations, 212 compounds with a defined activity value in the METI database were used as external test set, excluding those already present in the NCTRER database. Also for this test set, chemical

descriptors were calculated on the basis of the procedure described in paragraph 4.1.2 and the model equation was implemented in MS Excel.

Model 2

To validate Model 2 some data were selected from a paper by Sutherland *et al.* [74] who studied a compilation of ER ligands from NTP (ER-tox set). This compilation of binding affinities for 638 substances was prepared by the National Toxicology Program at the National Institute of Environmental Health Sciences, with the purpose of evaluating the performance of various *in vitro* ER binding assays¹.

It was found that there was reasonable correspondence between binding affinities determined by the various assays. From this collection, the authors selected 616 single chemical substances that were non-redundant. If compounds were tested by multiple assays, an average RBA value was calculated, excluding assays that used ER β or ER from non-mammalian species. The latter sources were excluded because they are reported for few substances, making the evaluation of their compatibility with other assays more difficult. Due to the heterogeneity of the experimental data sources the quality of these data was judged too poor for a quantitative analysis but satisfactory for a classification analysis. The activity classes and the structures were provided within the supporting information. Among the 616 substances 225 were in common with the NCTRER dataset. For them, the activity class was compared in the two datasets using the presence or absence of any detectable activity. Results are shown in Table 4.2.

A few compounds that were inactive in the NCTRER dataset, had slight activity in the NTP compilation, with an activity range of 0.00004-0.007 for RBA. For this reason, in order to have a higher consistency between the two data sources a

¹ http://iccvam.niehs.nih.gov/docs/endo_docs/final1002/erbdndbrd/ERBd034504.pdf

limited number of compounds having an activity in the NTP dataset within this range were excluded from the analysis.

Table 4.2 Comparison of the activity class assigned to compounds in common in the NTP compilation, as reported by Sutherland *et al.* [74], and NCTRER dataset.

		NCTRER	
		Active	Inactive
NTP	Active	127	16
	Inactive	0	82

The final test set extracted from NTP resulted in 368 compounds (124 inactive and 244 active). In order to assign the activity class to the test set to decide whether or not a compound was active on the basis of Figure 4.1, the workflow procedure was applied by visual inspection of the compound structures. In particular no special difficulties were raised in assigning features F1, F2, F3, F4 and F5, because the requirements (a ring, a phenolic ring, presence of heteroatoms) were very simply detected. More problematic was the assignment for F6, since several aspects (logP, H bonding, rigid structure, etc...) had to be taken into account contemporaneously. It also implies a more subjective assignment of this feature by balancing several considerations. In order to make this assignment as objective as possible a specific strategy was adopted. All compounds were grouped based on their chemical classes and then within each class specific reasons for F6 assignment were investigated on the basis of compounds belonging to the NCTRER dataset. By examining the characteristics for F6 assignment in each chemical class for the training compounds a rationale for extending the F6 value to all the other compounds was derived. The complete list of assigned features is available in Annex A.

Model 3

In the original publication [73] a second data set, designated as ER1092, was also investigated by the authors. It is an aggregation of data from the literature containing 1092 chemicals, of which 350 are active and 736 are inactive. For this large dataset the authors provided Molconn-Z descriptors. After eliminating the compounds already contained in the NCTRER dataset, 860 compounds were used as external set (225 active and 635 inactive). For this specific model no further analysis of the 2D descriptors originally calculated by the authors was done and the file provided by the authors was used directly.

4.1.4. APPLICABILITY DOMAIN ASSESSMENT

Another important issue relates to the definition of the applicability domain (AD) and possible solutions for defining it. Different concepts and methods to define the AD have been applied and tested on the external sets used to validate the models, in order to verify their utility in increasing the confidence on the predicted results, and to determine the boundaries for the validity of the models.

Principal Component Analysis was used to represent training and test sets and to compare visually their relative distribution. To calculate the principal components, the descriptors used in developing the model were normally selected, except for Model 2, where the six features were not descriptive enough. In this case a restricted pool of Dragon descriptors [121] were instead adopted. Only some classes of 2D descriptors were computed (constitutional, topological, walk and path counts, connectivity and information indexes, topological charge indexes and few molecular properties). Constant and near constant descriptors and descriptor pairs correlated with $r > 0.95$ were excluded and consequently 125 descriptors were used for the PCA analysis.

To explore more completely the AD definition a program specifically designed for this purpose was tested. AMBIT Disclosure [122] is a program implementing several methods to assess similarity between a group of training compounds and test set

compounds. They include both structure-based methods, based on a predefined set of fingerprints and fragments, and descriptor-based methods, where descriptors can be imported from other sources. Different thresholds can be tested to focus more precisely the training set space.

Finally, in the case of Model 3 there is a probability level associated with each prediction and the span of this probability was used to assign a reliability measure to the results.

4.2. RESULTS AND DISCUSSION

4.2.1. MODEL 1

4.2.1.1. Internal validation

Model 1 was recalculated with Statistica [114] using the original descriptors and its validity and the associated statistical parameters were confirmed.

4.2.1.2. Sensitivity study

A few differences were noticed in the recalculated descriptors, due probably to the software used and the procedure adopted. As expected, 3D parameters were found to be sensitive to the optimisation procedure, reducing their reproducibility. In the diagrams below the recalculated descriptors for E_{torsion} (Figure 4.3) and E_{HOMO} (Figure 4.4) are compared. A very good reproducibility was obtained for E_{HOMO} while lower consistency was found for E_{torsion} .

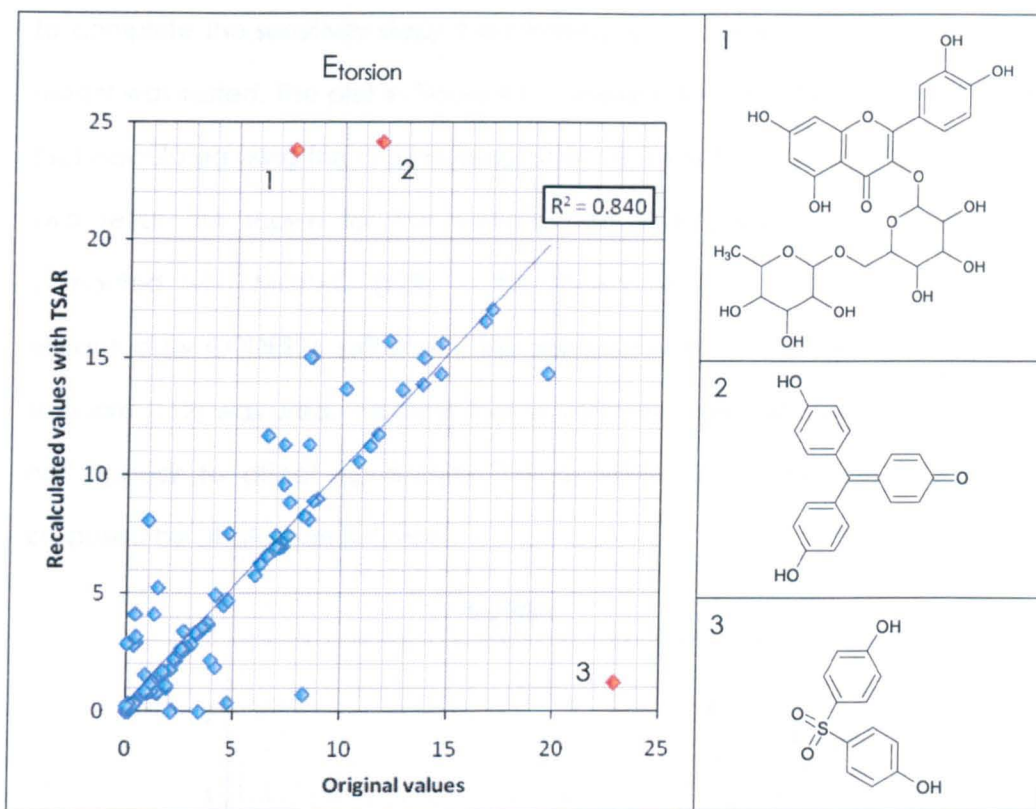


Figure 4.3 Correlation for the values of "Energy of torsion" descriptor. Three compounds in red in the plot are outliers and were not considered to derive the correlation coefficient.

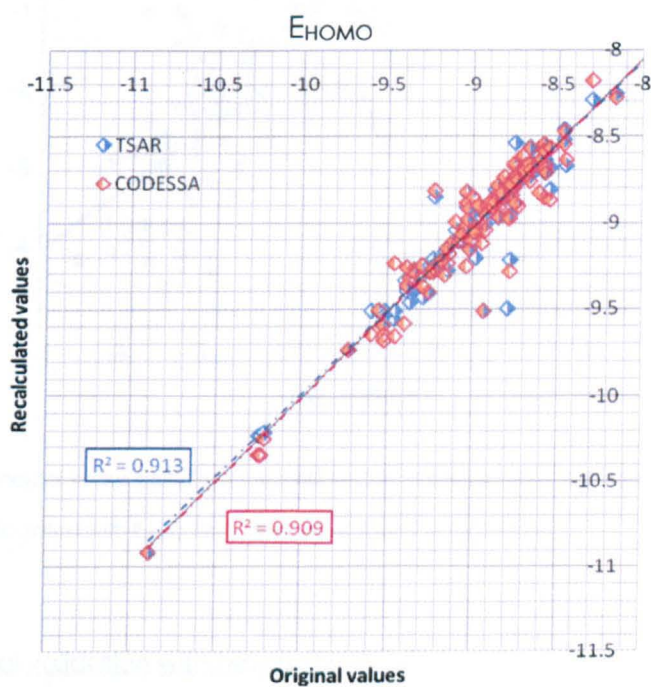


Figure 4.4 Correlation for the values of "HOMO Energy" descriptor calculated with TSAR or CODESSA with AM1 parameterization.

To complete the sensitivity study the influence of the descriptor variation on the model was tested. The plot in Figure 4.5 compares the experimental activity with that calculated using the original descriptors, or the set of newly computed ones. Two series are shown for the recalculated descriptors: in one case E_{HOMO} computed with MOPAC [123] in the OpenMolGrid environment [119] and extracted by CODESSA software [118] was used, while in the second one TSAR software [112] was used. It is clear that uncertainty in the descriptors values does not appear to affect significantly the activities calculated with the model proposed by Ghafourian & Cronin.

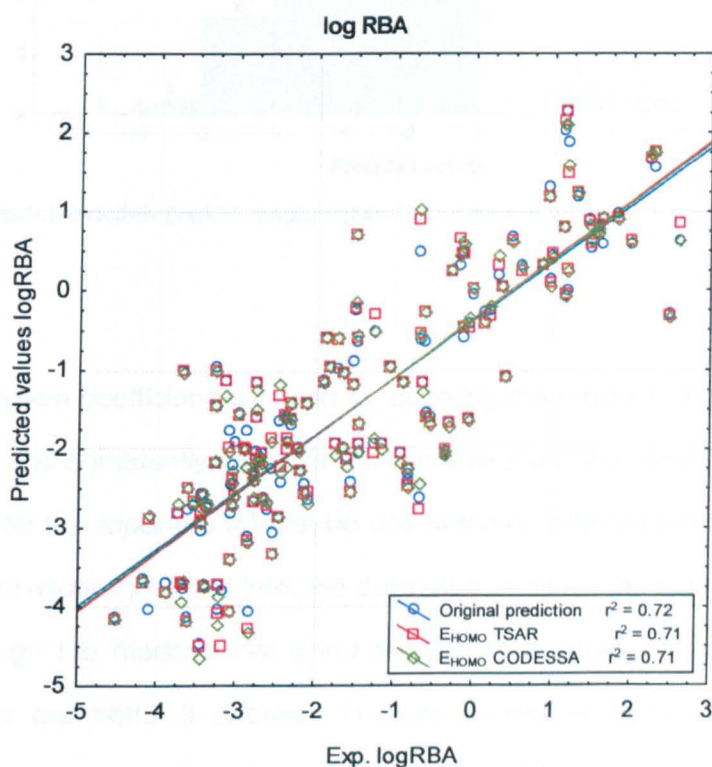


Figure 4.5 Experimental versus calculated activity for the model equations using the original descriptor set or those recalculated with E_{HOMO} from alternatively TSAR or CODESSA.

4.2.1.3. External validation with new test sets

For the inactive compounds of the NCTRER dataset, constituting the first test set, the predicted activity is reported in the histograms in Figure 4.6. The majority of

compounds (78%) are predicted with a relatively low activity ($\log\text{RBA} < -2$), a further 19% fall in the range from -2 to 0 and three are predicted highly active.

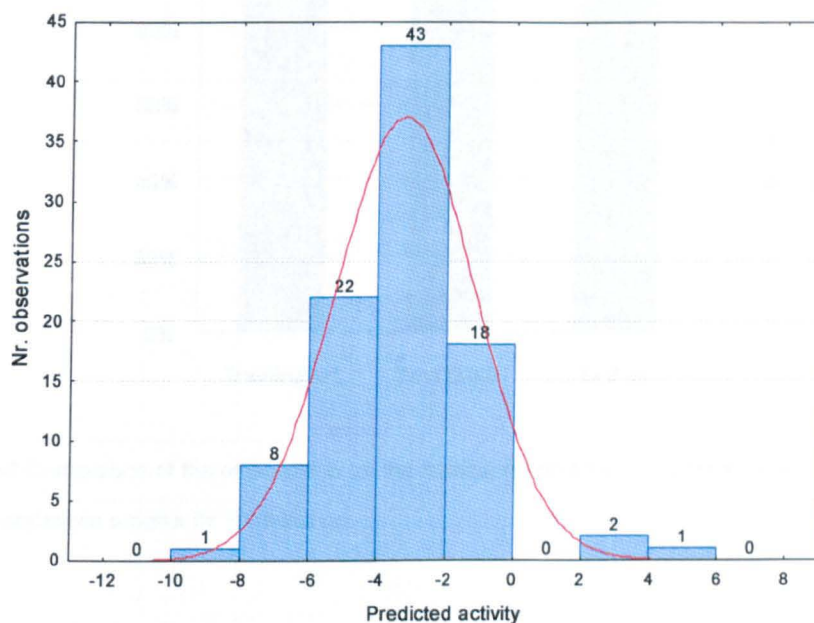


Figure 4.6 Predicted activity ($\log\text{RBA}$ range) distribution for the test set of inactive chemicals of NCTRER dataset.

The correlation coefficient obtained by applying the model to the EDKB set was $R^2 = 0.48$ and consistently lower for the Japanese set. The diminished accuracy obtained for the Japanese set can be due to the activity data of the second test set that are relatively diverse from the data used to derive the model.

Even though the model is not good enough to capture the data trend, the predictions are better if analysed in more qualitative terms. In Figure 4.7 a graphical comparison of the residuals between training and test sets is reported. Overall residuals for the predictions in both the test sets are within two log units for about 80% of the compounds; approximately 9% of compounds have residuals larger than 3 log units in both EDKB and Japanese test sets.

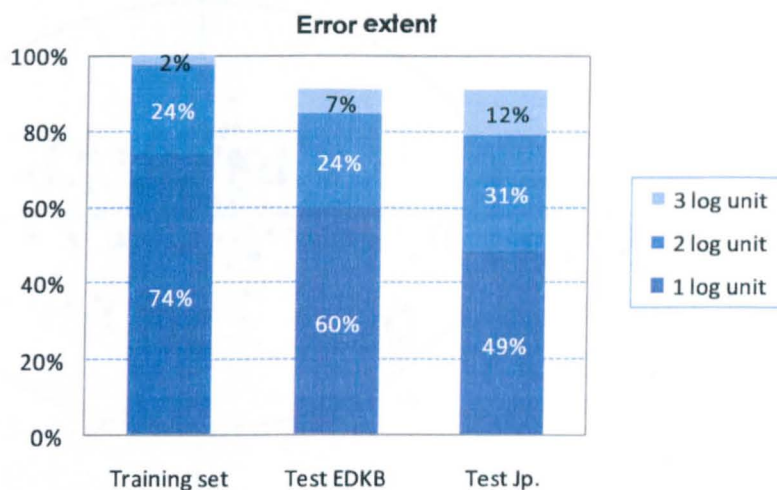


Figure 4.7 Comparison of the error extents for the training set and the two external test sets. Prediction performances are superior for EDKB test set.

4.2.1.4. Applicability domain analysis

AD was evaluated by exploring different possible definitions for outliers. In the case of inactive chemicals belonging to the NCTER dataset the main target was to identify possible reasons to account for the three compounds predicted highly active. The analysis was performed in terms of similarity assessment based on descriptor range (Table 4.3) or PCA score scatter plot (Figure 4.8). No special trend appeared to justify these three wrong predictions but the PCA score scatter plot highlights a differential distribution of active and inactive compounds.

Table 4.3 Descriptor range and outlier descriptor values.

Name	Min-Max range	Sitosterol	Cholesterol	4,4'-Methylenebis (2,6-di- <i>t</i> -butylphenol)
N _C	7 → 34	29	27	29
N _{halogens}	0 → 10	0	0	0
I _{phenol}	0 → 1	0	0	1
W	62 → 7474	2463	2022	2524
³ K _a	0.64 → 8.83	3.88	3.56	5.79
⁶ χ _{ch}	0.06 → 0.31	0.16	0.16	0.11
E _{torsion}	0.000114 → 22.94	8.41	15.67	10.02
E _{HOMO}	-10.89 → -8.16	-9.34	-9.35	-8.61

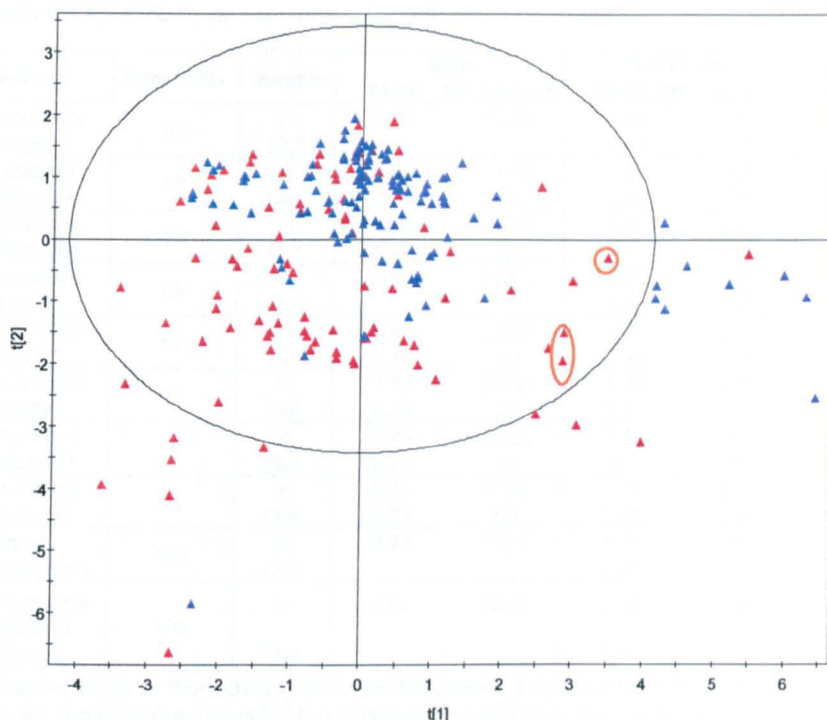


Figure 4.8 PCA score scatter plot for the first two PC (explained variance: 59.4%), calculated on the basis of the eight descriptors, for the training set (active compounds, blue triangles) and the test set of inactive compounds (red triangles). Three outliers are indicated by red circles.

For the other two external test sets, AMBIT Disclosure and the similarity measures there implemented were used. In Table 4.4 the RMSE for EDKB and Japanese sets is reported for compounds included or excluded in the applicability domain. Some of the methods (e.g. probability density distribution) seem more effective in separating compounds with a lower RMSE (i.e. with a good predicted value) from those having a higher RMSE.

A further analysis was done to determine whether, by using only compounds within the applicability domain, the percentage of compounds predicted within one or two log units increases. In Figure 4.9 and Figure 4.10 the results are shown: no substantial increase in these percentages was observed for both test sets.

Table 4.4 Applicability domain evaluation with AMBIT Disclosure.

Method	Threshold+	Domain	Train RMSE (N° comp.)	Test ER1092 RMSE (N° comp.)	Test Japan RMSE (N° comp.)
Descriptor Range*	100	In Out	0.93 (131) -	1.45 (32) 1.68 (13)	1.5 (164) 2.41 (48)
Euclidean Distance*	99	In Out	0.93 (130) 0.76 (1)	1.59 (41) 0.44 (4)	1.73 (202) 2.11 (10)
Euclidean Distance*	95	In Out	0.95 (124) 0.77 (7)	1.39 (35) 1.91 (10)	1.67 (180) 2.14 (32)
Probability Density*	100	In Out	0.93 (131) -	1.31 (38) 2.35 (7)	1.63 (189) 2.54 (23)
Probability Density*	99	In Out	0.94 (130) 0.76 (1)	1.31 (38) 2.35 (7)	1.62 (187) 2.52 (25)
Probability Density*	95	In Out	0.95 (124) 0.76 (7)	1.06 (29) 2.11 (16)	1.47 (158) 2.38 (54)
City-block Distance*	99	In Out	0.93 (130) 0.76 (1)	1.44 (41) 2.2 (4)	1.71 (204) 2.5 (8)
City-block Distance*	95	In Out	0.95 (124) 0.73 (7)	1.33 (31) 1.88 (14)	1.63 (161) 2.09 (51)
Missing Fingerprints	100	In Out	0.93 (131) -	1.31 (24) 1.73 (21)	1.42 (104) 2.03 (108)
Fingerprints Tanimoto distance	100	In Out	0.94 (131) -	1.44 (38) 1.91 (7)	1.73 (202) 2.06 (10)

+ Percentage of the training set used for deriving applicability domain.

* Data standardised and PCA performed for pre-processing data.

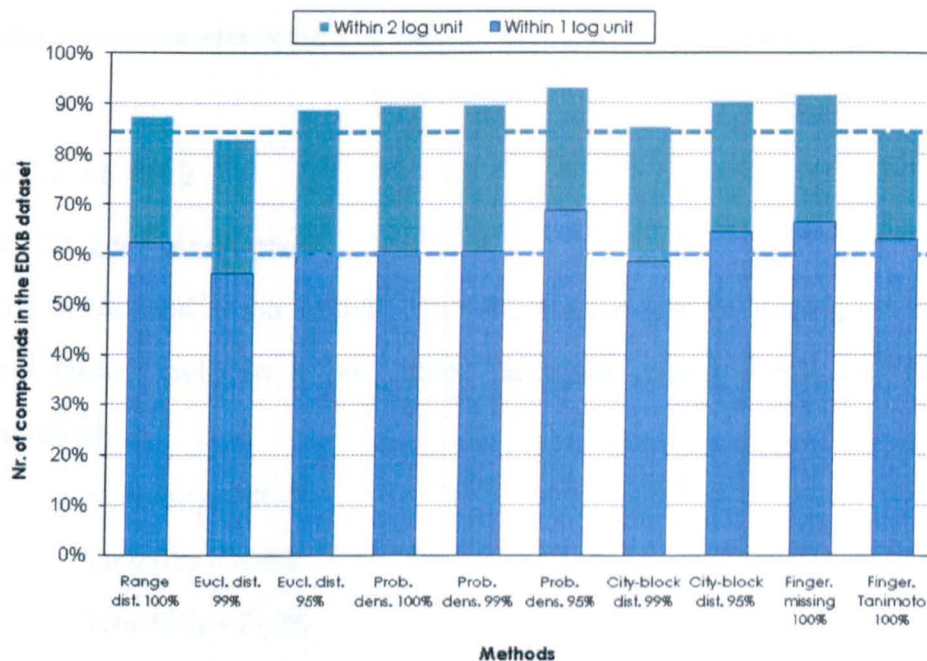


Figure 4.9 Percentage of compounds with residuals within one of two Log units for the first test set. Different ways for estimating applicability domain are compared with the reference values (pale blue and turquoise green dotted lines) where all compounds belonging to the test set are used.

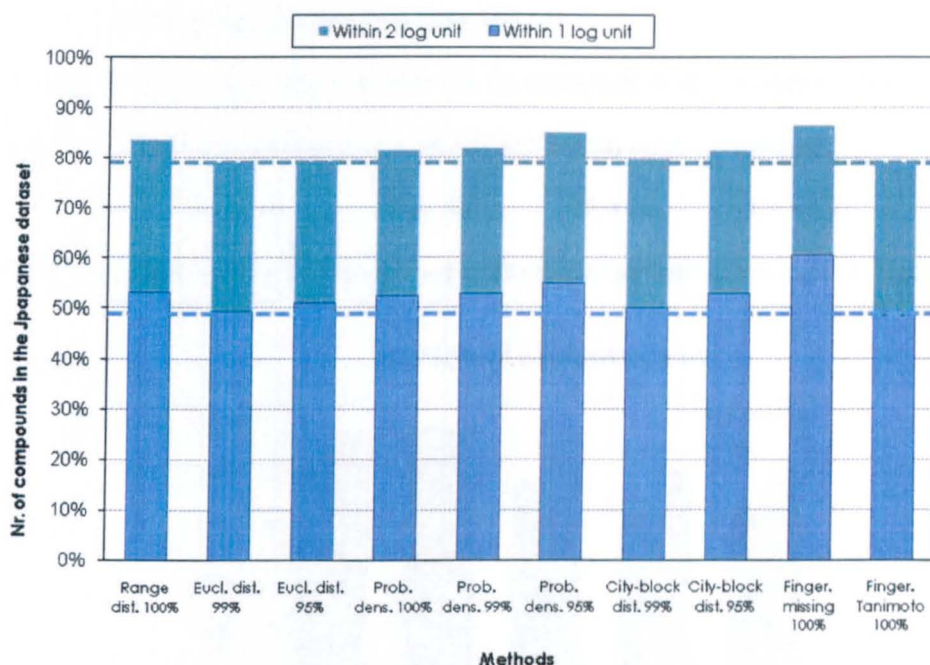


Figure 4.10 Percentage of compounds with residuals within one of two Log units for the second test set. Different ways for estimating applicability domain are compared with the reference values (pale blue and turquoise green dotted lines) where all compounds belonging to the test set are used.

4.2.2. MODEL 2

4.2.2.1. Internal validation

The application of the flowchart to the compounds in the training set, by using the feature assigned in the DSSTox database gave the following statistical parameters:

Accuracy = 82.3%

Sensitivity = 90.8%

Specificity = 71.3%

MCC = 0.64

S = 63.3%

4.2.2.2. External validation with a new test set

By applying the flowchart to test set compounds and comparing the assigned activity with that available in the ER-tox set, results were as follows:

TP = 218 TN = 108 FP = 16 FN = 26

Statistical parameters for training set and the test set are compared in Figure 4.11.

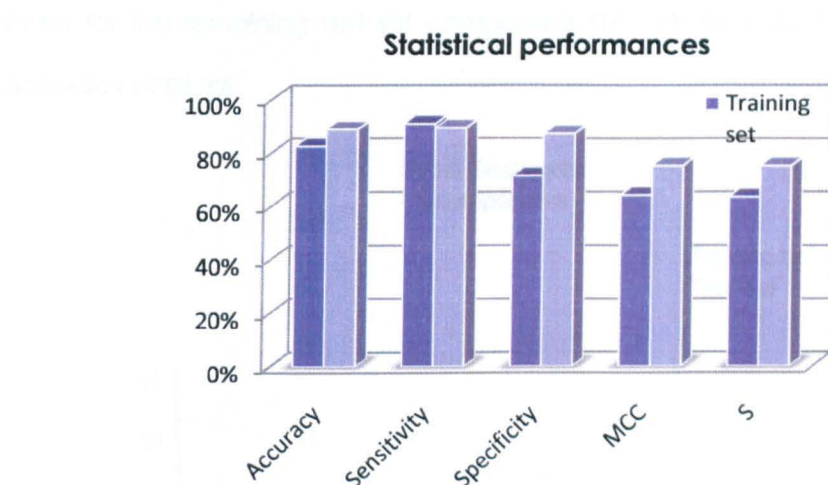


Figure 4.11 Comparison of the classification performances for the training and the test set.

The model seems quite stable, performing on the test set even better than on the training set. The main reason for this good performances can be based on the limited structural complexity of compounds belonging to the ER-tox set: a lot of chemicals belong to few chemical classes (PCBs, PAHs, Steroids, Stilbenes and Triphenylethylenes) so that for chemical classes well represented both in the training and in the test set the predictions are very good and this can affect the overall resulting performances.

At the same time a main drawback of this model is the subjective way of assigning the very complex feature F6.

4.2.2.3. Applicability domain analysis

Dragon 2D descriptors were used to represent training and test set in a PCA analysis (Figure 4.12).

A group of 30 compounds are outside the Hotelling ellipse (significance level = 0.05) – figure not shown – but the predictions for them are reasonably similar to those for the remaining test set compounds (TP = 3, TN = 22; FN = 5) with an accuracy of 83.3%.

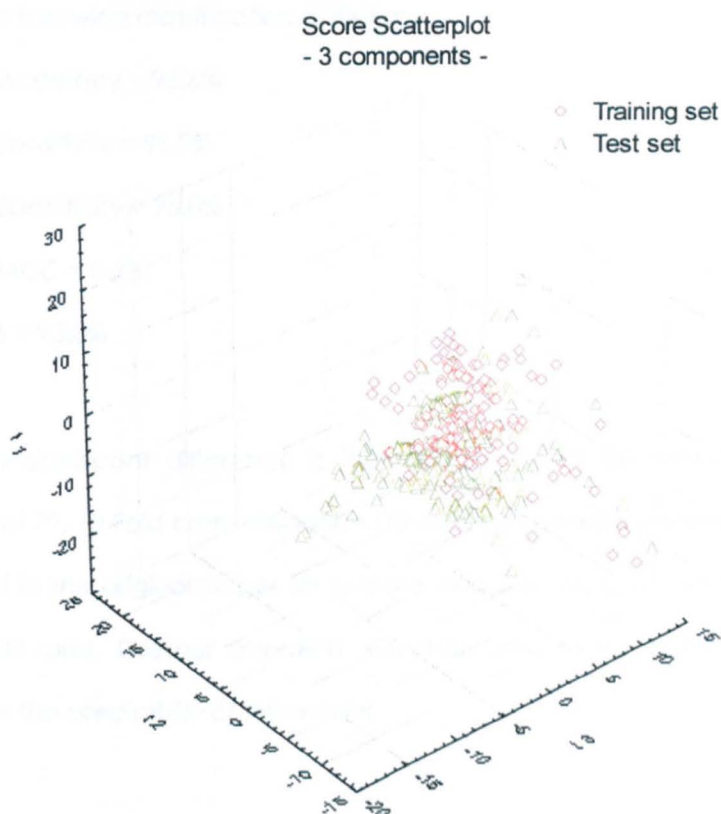


Figure 4.12 PCA score scatter plot for the training and test set using the first three components (explained variance = 47.2%).

4.2.3. MODEL 3

4.2.3.1. Internal validation

The model was re-developed starting from the original dataset provided as txt file and using the default options in the DF program. The model is a combination of six individual trees, giving comparable results with those reported in the original publication:

TP = 127 TN = 97 FP = 4 FN = 4

With the following classification statistics:

Accuracy = 96.6%

Sensitivity = 96.9%

Specificity = 96.0%

MCC = 0.93

S = 93.0%

The only significant difference is the selection of 84 descriptors in the model instead of 79. 10-Fold cross-validation (10 runs) gave similar performances to those reported in the original paper for a more extensive cross-validation exercise (10-fold, 2000 runs). Greater attention was then paid to an external validation to evaluate the predictivity of the model.

4.2.3.2. External validation with a new test set

By applying the DF model to predict the ER1092 test set the classification results were:

TP = 161 TN = 420 FP = 215 FN = 64

Statistical parameters for the test set were:

Accuracy = 67.6%

Sensitivity = 71.6%

Specificity = 66.1%

$$\text{MCC} = 0.33$$

$$S = 31.0\%$$

Performances are considerably lower than those on the training set, indicating the risk of an overfitted model. To judge better about this possibility, the influence of chemical structural diversity was analysed through the AD assessment.

4.2.3.3. Applicability domain analysis

Several approaches for evaluating the applicability domain were used:

1) A PCA analysis on the 84 selected descriptors was performed and reported in Figure 4.13 but from this plot is it difficult to draw conclusions about representativeness of the test set.

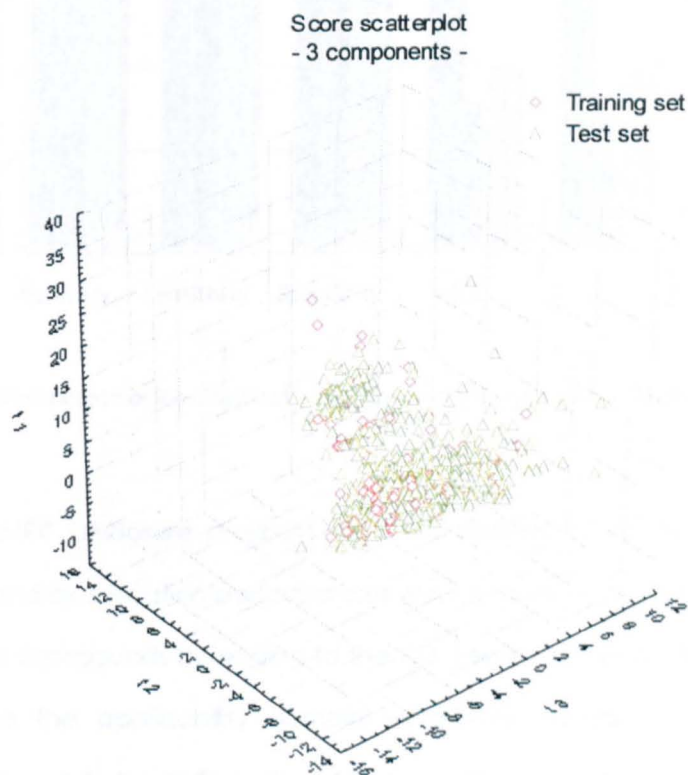


Figure 4.13 PCA score scatter plot for the training and test set using the first three components (explained variance = 56.2 %).

2) In the original paper it was suggested that by using only those compounds with a higher level of confidence, performances increased. This concept was applied by using a subset of the test set excluding compounds with probability in the class assignment between 0.3-0.7. In Figure 4.14 the results for these two subsets are compared with the entire test set and a slightly improvement is shown for the latter but this is not really significant. At the same time this increased accuracy implies a reduced number of compounds to be considered in the AD (178 compounds excluded).

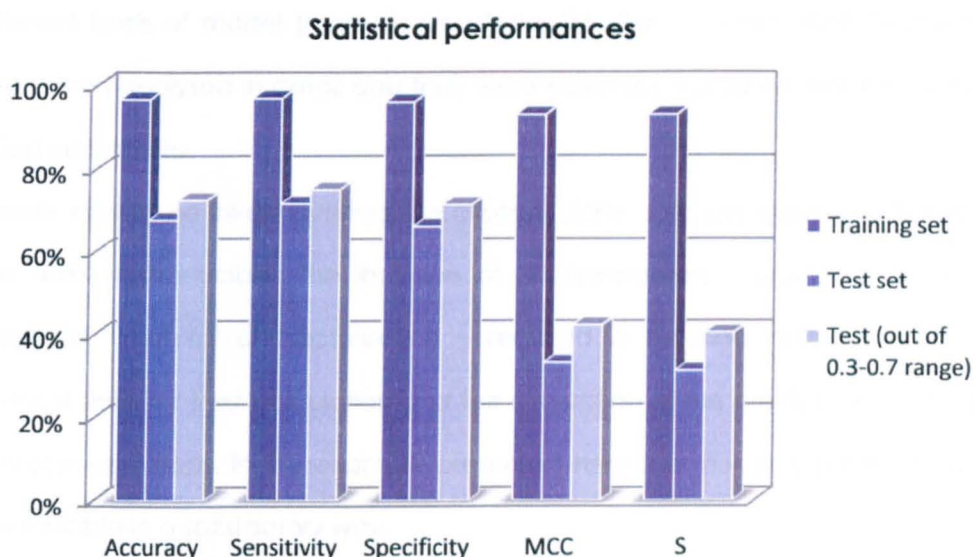


Figure 4.14 Comparison of the classification performances for the training and the test set.

3) The AMBIT Disclosure program was also applied to define the applicability domain and classification performances were evaluated as shown in Table 4.5 by using only compounds belonging to the AD. Performances for the compounds in or outside the applicability domain were very variable depending on the technique used to define the AD but still none produced a significant improvement.

Table 4.5 Classification performances for the compounds belonging to the test set considered in or outside the AD.

	Total	Range T100 in	Range T100 out	Euclidean T95 in	Euclidean T95 out	City-block T95 in	City-block T95 out	Probability T100 in	Probability T100 out
Accuracy	67.6	71.0	66.8	67.0	75.4	68.5	61.7	67.3	67.7
Sensitivity	71.6	83.3	66.7	70.8	100.0	73.6	54.2	81.5	64.7
Specificity	66.1	62.5	66.8	65.6	72.5	66.6	63.5	60.3	68.6
MCC	33.4	45.4	28.5	32.6	46.7	36.0	14.5	39.4	28.7
S	31.0	43.3	26.2	30.4	35.7	33.6	12.9	36.0	26.7

4.3. CONCLUSIONS

Different kinds of model (regression and classification, SAR and QSAR models) have been analysed in detail and they were externally validated with new data found in literature.

Models relying on bi-dimensional descriptors alone seemed more user-friendly and more reproducible. The inclusion of 3D parameters – providing a more complete structural characterisation – required a detailed definition of the protocol used for their calculation and the evaluation of the prediction sensitivity to procedural steps. In the example presented here the model was robust and reproducible in a satisfactory way.

The linear model was able in prediction to detect the activity range for a substance but did not completely catch the activity trend. Concerning the hypothesis of using computational models as pre-screening tools for large inventories of chemicals, - many of them probably inactive - SAR or classification models may be preferable.

To avoid overfitting especially when the model complexity is increasing, an accurate validation is essential, including an external test set.

Several ways for assessing the applicability domain have been evaluated depending on the available information. Although some methods seem better than others no efficient way to detect poor predictions has been identified.

PART III

DEVELOPMENT OF NEW QSARs FOR OESTROGENICITY

CHAPTER 5

BINARY CLASSIFICATION MODELS FOR SCREENING HETEROGENEOUS DATASETS FOR THEIR OESTROGENIC ACTIVITY

5.1. INTRODUCTION

Based on the analysis conducted in Chapter 4 the experimental setting for developing new models for oestrogenicity was decided. Preference was given to classification models since the modelling performances achievable so far can be considered more qualitative than quantitative, due probably to data quality. Possibly the classification approach here adopted can be coupled with quantitative models.

The most efficient solution for providing a prioritisation tool is to develop simple models, easy to use, so preference was given to the use of 2D descriptors. Moreover, it has already been observed that often similar performances can be achieved with the use of 2D and 3D descriptors [62].

In this way it was possible also to rely on a large dataset, close to a thousand heterogeneous compounds, to derive robust models accurately validated. The focus was given to multiple endpoints the better to characterize the effects of EDs evaluating both binding and transcriptional activity.

5.2. MATERIAL AND METHODS

5.2.1. DATASET

Activity data

As a source of activity data the Japanese METI database was used [120]. Previous studies showed that these data are in relatively good agreement with other databases (see Paragraph 4.1.3). This database is one of the largest collections of data for ER publicly available, with more than 900 compounds. It contains experimentally determined values of human ER alpha for both receptor binding (RBA) and reporter gene (RA) assays expressed as molar percentage of activity using 17 β -estradiol as reference. To develop binary classification models any detectable activity in the test was associated with the "active" class while those compounds with no detectable activity were labelled "inactive". The dataset is reported in Annex B. It represents a heterogeneous dataset of compounds, including natural and synthetic steroids, drugs and chemical contaminants such as pesticides, PCBs and phthalates.

Chemical structures and descriptors

Chemical structures were sketched and, for salts, the free acid or basis form was used. Adopting a very simple approach only a 2D configuration of the molecules was used while the 3D conformation and stereo configuration were ignored. For this reason 2D duplicates of different 3D isomers were included only once, verifying that the associated activity class was comparable for all possible forms sharing the same 2D structure (only 2 structures did not satisfy this requirement and were discarded). A further 100 compounds whose RBA values were not determined were excluded to consider both the endpoints on a similar basis. The final dataset comprised 806 single 2D structures, with the majority of the compounds considered inactive (see Figure 5.1).

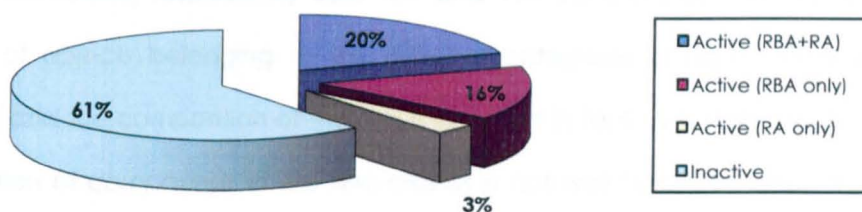


Figure 5.1 Compounds distribution in the classes of activity for RBA and RA.

Dragon software [121] was used to compute 929 2D descriptors. This number was reduced by excluding constant and nearly constant variables (diverse for less than 6% of compounds), pairwise correlated variables (with a K correlation greater than 0.95) and those with missing values, to reduce redundant and useless information. Finally a total of 250 descriptors were retained and submitted to the autoscaling procedure.

The dataset so collected was split into three parts: the training set constituted by the examples provided to the learning algorithm, the validation set used to assess which are the best parameters and architecture for the models, and a test set to assess independently a group of compounds never used to validate the model's performances in prediction. To perform this selection an unsupervised learning method based on a Kohonen map, also called self-organizing map (SOM), was used. This type of NN is trained to produce a bi-dimensional representation of the input space in a map, preserving its topological properties. This makes SOM useful for visualizing high-dimensional data. After the training the input samples are located in the map in the neurons on the basis of the similarity of their input vectors.

A Kohonen map of 15x15 neurons dimension was trained for 500 epochs using the descriptor matrix previously prepared. On the basis of the distribution of the

compounds in the top map the data set was split into training, validation and test sets, each containing respectively 506, 150 and 150 compounds. The correct proportion of objects belonging to the different categories in Figure 5.1 was maintained and the composition of the sets is reported in Table 5.1. Unfortunately the distribution of compounds in the two classes is not well balanced since the majority are inactive for both end points.

Table 5.1 Compounds repartition in training validation and test sets.

	RBA		RA	
	Active	Inactive	Active	Inactive
TRAINING SET	180	326	117	389
VALIDATION SET	54	96	35	115
TEST SET	54	96	35	115

5.2.2. MODELLING METHODS

Different modelling methods were investigated, ranging from simpler and more intuitive models such as classification trees to more sophisticated architectures, including NN.

5.2.2.1. Classification and Regression Tree (CART)

The classic CART algorithm, developed by Breiman *et al.* [124,125], uses the methodology of tree building as a hierarchical classification method. The purpose of this analysis is to determine a set of if-then logical (split) conditions that permit accurate classification and prediction of cases that are easy to interpret, yet it takes into account the fact that different relationships may hold among variables in different parts of the data. CART formulates simple if-then rules for binary recursive partitioning of all the objects into smaller subgroups, where the compounds belonging to the dataset, represented by a "node" in a decision tree, can be split into only two groups. Thus, each node can be split into two new

"branches". The goal of this process is to maximize homogeneity of the values of the dependent variable Y in the various subgroups. The CART technique is essentially non-parametric, and does not rely on any particular assumptions about the type of dependence of the dependent variable Y on predictors X_i (in contrast to various regression techniques) or about statistical properties of the data. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure.

CART is scale-invariant, extremely robust with respect to outliers, and does automatic stepwise variable selection. It performs well when the pattern space can be separated into pure class subspaces.

The process of computing classification trees can be characterised as involving four basic steps:

1. *Specifying the criteria for predictive accuracy.* The most accurate prediction is defined as the prediction with the minimum costs, an extension of the misclassification rate modulated by the hazard associated with wrong predictions. The cost for the errors was considered equally important for the two classes while the *a priori* probability for the two classes was set as equal so that the relative size of the prior assignments to each class can be used to "adjust" the importance of misclassifications for each class.

2. *Selecting splits.* In general terms, the best split at each node is determined in terms of the reduction in impurity (heterogeneity) giving the greatest improvement in predictive accuracy. All the splits are ranked and the variable that achieves the highest purity at root is selected. The exhaustive search for univariate splits method works by searching for the split that maximizes the reduction in the value of the selected goodness of fit measure. This is usually

measured with some type of node impurity measure, which provides an indication of the relative homogeneity of cases in the terminal nodes. The Gini measure of node impurity was used. It is a measure which reaches a value of zero when only one class is present at a node.

Each node is assigned to a predicted class based on the following criteria:

$C(j|i)$ is the cost of classifying i as j ;

$\pi(i)$ is the prior probability of i ;

N_i is the number of class i in the dataset;

$N_i(t)$ is the number of class i in a node;

Node is class i , if:

$$\frac{C(j|i)\pi(i)N_i(t)}{C(i|j)\pi(j)N_j(t)} > \frac{N_i}{N_j}$$

for all values of j .

3. *Determining when to stop splitting.* The tree is allowed to grow until all terminal nodes are pure or contain no more misclassified cases than a specified minimum fraction of objects.

4. *Selecting the "right-sized" tree.* The "right-sized" tree should be sufficiently predictive, but at the same time it should be as simple as possible. It should exploit information that increases predictive accuracy and ignore information that does not. The performances on the validation set were used to select efficiently the well-dimensioned tree; then the best tree was evaluated on the test set.

The software implementing the CART algorithm used in this study was Statistica [114].

5.2.2.2. Decision Forest (DF)

As an extension of normal classification trees, DF was also used since it has already been proposed as an algorithm for modelling oestrogenicity. The basic idea of the method has already been introduced in Chapter 4 and it is described in more detail elsewhere in the literature [126].

DF is a consensus modelling technique that combines multiple Decision Tree models, maximizing in their construction the use of diverse descriptors. Different parameters can be optimised including the minimum and maximum number of trees, and the minimum number of compounds allowed to enter in a node.

An application implementing the DF method was used that was available on the web¹.

5.2.2.3. Adaptive Fuzzy Partition (AFP)

To select, amidst the molecular descriptors series, the best parameters for classifying the data, a hybrid selection algorithm (HSA) based on Genetic Algorithm (GA) concepts was used, specifically designed to select relevant descriptors for classification aims [127]. GA methods, inspired by population genetics, consist of a population of individuals competing on a "survival of the fittest" basis. Each individual, or chromosome, represents a trial solution of the problem to solve. In the context of descriptor selection, the structure of the chromosome is very simple. Each descriptor is coded by a bit (0 or 1) and represents a component of the chromosome; 0 defines the absence of the descriptor, and 1 defines its presence. The algorithm, transforming the chromosome population into a new population with more adapted individuals, proceeds in successive steps called generations. During each generation, the population evolves by means of a "fitness" function that selects individuals by

¹ <http://www.fda.gov/nctr/science/centers/toxicoinformatics/DecisionForest/index.htm>

standard crossover and mutation operators. Crossover phase takes two chromosomes and produces two new individuals, by swapping segments of genetic material, i.e. bits in this case. Mutation randomly removes bits with a small probability of success from the chromosome population.

GA is very effective for exploratory search, applicable to problems where little information is available, but it is not particularly suitable for local search. Thus, a stepwise approach was combined with GA in order to reach local convergence, as it is quick and adapted to find solutions in "promising" areas already identified.

The index proposed as fitness function to evaluate the discrimination power of a selected subset of descriptors is based on a Fuzzy Clustering (FC) procedure [128]. A FC algorithm, where clusters are derived from fuzzy sets, can be considered as a generalisation of the traditional cluster procedure. These clusters are derived by assigning to each compound a number between 0 and 1, called degree of membership. A compound is defined by its degree of membership to each cluster, while a cluster can be characterised from the list of associated compounds with the highest membership degrees.

This index has the advantage that it can be quickly calculated and that one can also estimate the descriptor relevance by analysing complex molecular distributions, in which finding boundaries between the different categories is difficult.

To prevent over-fitting and a poor generalisation, a cross-validation procedure was included in the algorithm during the selection procedure, randomly dividing the database into training and validation sets. The fitness score of each chromosome was derived from the combination of the scores of the training and validation sets. More details about the HSA procedure and the proprietary software used can be found in Ros *et al.* [127].

The following parameters were used in the data processing:

(i) fuzzy parameters: weighting coefficient = 1.5; tolerance convergence = 0.01; number of iterations = 100; cluster number = 10.

(ii) genetic parameters: chromosome number = 20; chromosome size = total number of descriptors used; initial active descriptors in each chromosome = 8; crossover point number = 1; probability of selection = 0.5; probability of crossover = 0.5; probability of mutation = 0.1; probability of rejection = 0.2; number of generations = 10.

(iii) stepwise parameters: ascending coefficient = 0.02; descending coefficient = -0.02.

After the descriptor selection with HSA algorithm, the Adaptive Fuzzy Partitioning (AFP) method was used for classification purposes.

AFP is a supervised classification method implementing a fuzzy partition algorithm [129]. Fuzzy logic (FL) mimics human reasoning in its use of approximate information and uncertainty to generate decisions about intrinsically imprecise problems. The FL concepts indeed provide mathematical rules and functions able to calculate intermediate values between "absolutely true" and "absolutely false", called degrees of membership and ranging from 0 to 1.

It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces defined by fuzzy rules. The aim of the algorithm is then to select the descriptor and the cut position which allows one to achieve the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighbouring subspaces. Only the best cut is selected to divide the original subspace. For instance, in Figure 5.2, a graphical representation of fuzzy partitioning is presented. Three cuts per axis are tested from the original bi-dimensional descriptor space. As cut x_1 is the best, two subspaces are generated and considered to be further divided. Then, cut y_3 is

selected, but the procedure evaluates useful partitioning only in subspace S_2 ; finally, three subspaces are built.

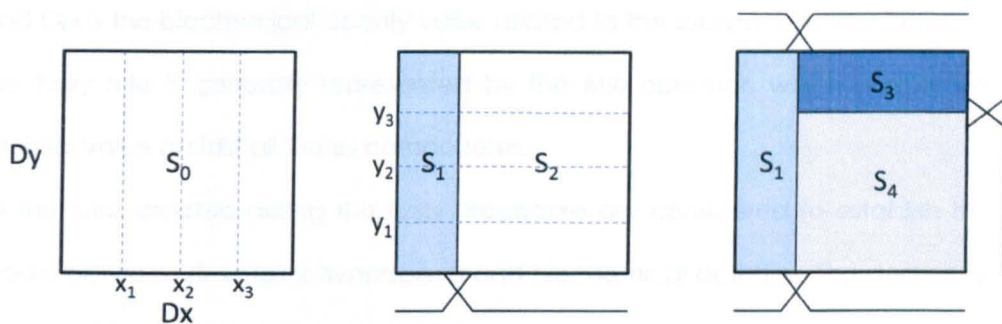


Figure 5.2 Example of adaptive fuzzy partitioning of a bi-dimensional space.

These techniques of rule generation are very simple as all the fuzzy rules can be formulated by linguistic labels. However, their performances in a database classification depend on the choice of partition selected. Generally, a coarse partition leads to a generalist system but also to a model where prediction results are too approximate; a fine partition leads to an accurate model of classification but also to a non-generalist system. To overcome this drawback, a fuzzy classification method was proposed, which simultaneously uses several fuzzy partitions of different sizes in a single fuzzy rule-based classification system. This approach allows one to obtain a good compromise between generalist and specialist systems, thereby improving the classification performances.

Assuming that the working space is a N -dimension hyperspace defined by N molecular descriptors, each dimension i can be partitioned into L intervals l_{ij} , where j represents an interval in the partition selected. Indicating with $P(x_1, x_2, \dots, x_N)$ a molecular vector in the hyperspace, a *rule* for a subspace S_k , derived by combining N intervals l_{ij} , is defined by:

"if x_1 is associated with $\mu_{1k}(x_1)$ **and** x_2 is associated with $\mu_{2k}(x_2)$...

and x_N is associated with $\mu_{Nk}(x_N) \Rightarrow$ the score of the activity O for P is O_{Kp} "

where x_i represents the value of the i^{th} descriptor for the molecule P , μ_{ik} is a trapezoidal membership function related to the descriptor i for the subspace k , and O_{kp} is the biochemical activity value related to the subspace S_k . The "and" of the fuzzy rule is generally represented by the *Min* operator, which selects the minimal value amidst all the μ_{ik} components.

All the rules created during the fuzzy procedure are considered to establish the model between descriptor hyperspace and biochemical activities. The degree of membership to the subspace S_k can be represented by:

$$O_k = \frac{\sum_{j=1}^M (\text{Min}_i^N \mu_{ik}(x_i)_{P_j} \cdot A_{P_j})}{\sum_{j=1}^M (\text{Min}_i^N \mu_{ik}(x_i)_{P_j})}$$

where M is the number of molecular vectors in a given subspace, N is the total number of descriptors, $\mu_{ik}(x_i)_{P_j}$ is the fuzzy membership function related to the descriptor i for the molecular vector P_j , and A_{P_j} is the experimental activity of the compound P_j . A classic procedure of centroid defuzzification is implemented to determine the chemical activity of a new test molecule. All the k subspaces are considered and the general formula to compute the degree of membership of the activity O for a generic molecule P_j is:

$$O(P_j) = \frac{\sum_{k=1}^{N_{\text{subsp}}} (\text{Min}_i^N \mu_{ik}(x_i)_{P_j} \cdot O_k)}{\sum_{k=1}^{N_{\text{subsp}}} (\text{Min}_i^N \mu_{ik}(x_i)_{P_j})}$$

where N_{subsp} represents the total number of subspaces.

The following AFP parameters were used to process the data set: maximal number of rules for each toxicity class = 30; range for the number of compounds for a given rule = 2-10; range of cuts for each axis = 2-10, p range: 1.05-1.6; q range: 0.4-0.95 (where p and q allow defining the trapezoidal membership function).

5.2.2.4. Multilayer perceptron (MLP)

A multilayer perceptron (MLP) is a feedforward network of neurons called perceptrons, introduced by Rosenblatt in 1958 [130].

The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then generating the output through some nonlinear activation function:

$$y = \varphi\left(\sum_{i=1}^n w_i x_i + b\right) = \varphi(\mathbf{w}^T \mathbf{x} + b)$$

where \mathbf{w} denotes the vector of weights, \mathbf{x} is the vector of inputs, b is the bias and φ is the activation function. The activation function chosen is the hyperbolic tangent $\tanh(x)$. This function is used because it is mathematically convenient and close to linear near the origin while saturating rather quickly away from the origin. This allows MLP networks to model well both strongly and mildly nonlinear mappings. While single-layer networks composed of parallel perceptrons are rather limited in what kind of mappings they can represent, the power of an MLP network with only one hidden layer is surprisingly high.

A typical MLP network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes between the input and output nodes, and an output layer of nodes. The interconnection matrix is restricted to feedforwarding activations (neither feedback nor self connections). Feedforward networks are instantaneous mappers; i.e. the output is valid immediately after the presentation of an input. The input signal propagates through the network layer by layer.

The computations performed by such a feedforward network with a single hidden layer with nonlinear activation functions and a linear output layer can be written mathematically as:

$$\mathbf{x} = \mathbf{f}(\mathbf{s}) = \mathbf{B} \varphi(\mathbf{A}\mathbf{s} + \mathbf{a}) + \mathbf{b}$$

where s is a vector of inputs and x a vector of outputs. \mathbf{A} is the matrix of weights of the first layer, a is the bias vector of the first layer. \mathbf{B} and b are, respectively, the weight matrix and the bias vector of the second layer. The function ϕ denotes the nonlinear activation function [131].

Figure 5.3 illustrates a simple MLP with one hidden layer. The circles are the neurons arranged in layers. The left column is the input layer, the middle column is the hidden layer, and the right column is the output layer. The lines represent weighted connections between neurons.

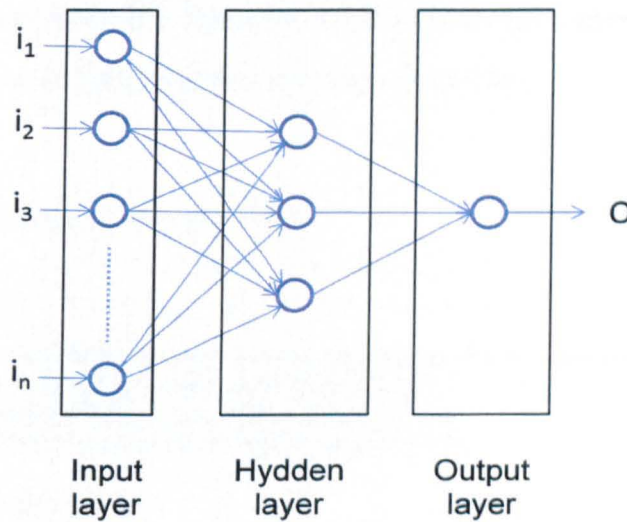


Figure 5.3 A MLP neural network structure. Perceptrons are arranged in layers.

MLP networks are typically used in supervised learning problems. This means that there is a training set of input-output pairs and the network must learn to model the dependency between them.

A neuron simply multiplies an input by a set of weights, and nonlinearly transforms the result into an output value. The power of neural computation comes from the massive interconnection among the neurons, and from the adaptive nature of the weights that interconnect the neurons. By adapting its weights, the neural network works towards an optimal solution based on a measurement of its

performance. For supervised learning, the performance is explicitly measured in terms of a desired signal and an error criterion.

The supervised learning problem of the MLP is solved with the back-propagation algorithm, which consists of different steps. In the forward pass, the predicted outputs corresponding to the given inputs are evaluated, with all connections feeding forward from inputs towards outputs. As a result the error between the desired output and actual output is computed. Then, in the backward pass, the error signal at the output units is propagated back through the network. Finally the synaptic weights and biases are updated using the results of the forward and backward passes using any gradient-based optimisation algorithm. The whole process is iterated until the weights have converged [131].

Table 5.2 Parameter setting used for GA and MLP.

GA parameters	
<u>GA setting</u>	
	Progression: generational (entire population is replaced with each iteration)
	Selection for the next generation: Roulette, rank-based
	crossover point number = 1, probability of crossover = 0.9
	Mutation operator: uniform, probability of mutation = 0.01
<u>GA training options</u>	
	Nr. Epochs: 1000
	Population size: 50
	Max generation nr.: 100
	Termination criteria: terminate after 250 epochs w/o improvement on validation set
	Class importance: use classes equally weighted
MLP parameters	
<u>Hidden Layers</u>	
	Nr. hidden layers: 1;
	Nr. perceptrons in the hidden layers: 4
	Transfer function: Tanh
	Learning rule: Momentum (Step size: 1, Momentum: 0.7)
<u>Output Layer</u>	
	Nr. perceptrons in the output layer: 2
	Transfer function: Tanh
	Learning rule: Momentum (Step size: 0.1, Momentum: 0.7)
<u>Supervised Learning Control</u>	
	Max nr. Epochs: 1000
	Termination criteria: Increasing in MSE on the validation set
	Weight update: batch

To derive the predictive models based on MLP, NeuroSolutions software was used [132]. Relevant descriptors were selected with GA in combination with MLP by optimizing as fitness function the costs on the validation set. The specific parameters used for GA are reported in Table 5.2 together with those used for MLP.

5.2.2.5. Support Vector Machine (SVM)

The support vector machine (SVM) is a classifier searching for a decision boundary - a hyperplane - that discriminates between the two classes [133]. In particular, as shown in Figure 5.4, SVM is designed to find the hyperplane with the largest distance to the closest points from the two classes, the maximal margin classifier. This is motivated by the concept of using only those inputs that are near the decision surface since they provide more information about the classification [134].

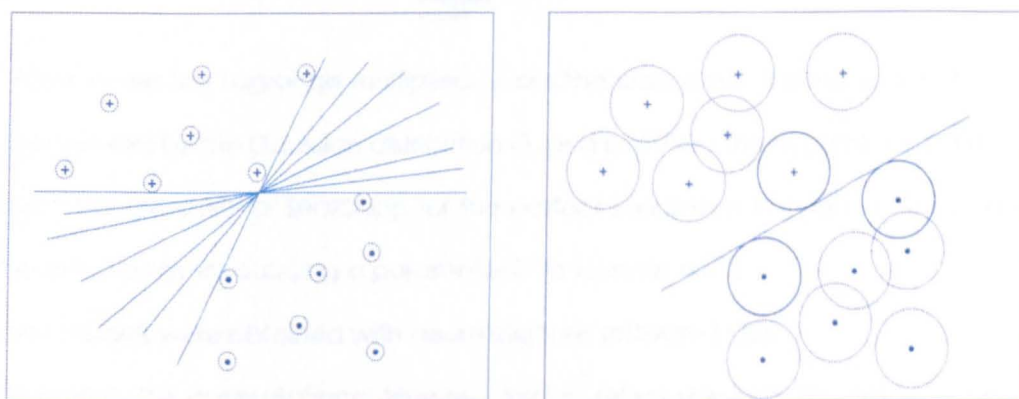


Figure 5.4 Exemplification of the SVM process for identifying the support vectors and the maximal margin classifier hyperplane. Adapted from [134].

A limitation of this approach is that in many cases classes cannot be separated by a hyperplane and a non-linear decision surface is required. This problem can be addressed with SVM by mapping the data from the original input space into a

feature space where a linear separator can be found. In this case SVM transforms the data into a high-dimensional space so that it can transform complex problems (with complex decision surfaces) into simpler problems that can use linear discriminant functions.

This mapping is obtained through a kernel function, which can be viewed as a distance between samples in feature space. In this study, this is done using as kernel a Radial Basis Function (RBF) network that places a Gaussian at each data sample. Thus, the feature space becomes as large as the number of samples. The RBF, however, uses backpropagation to train a linear combination of the Gaussians to produce the final result. The learning algorithm is based on the Adatron algorithm extended to the RBF network. After adaptation only some of the α_i are different from zero (the so called support vectors). They correspond to the samples that are closest to the boundary between classes. The output for testing is given by:

$$f(x) = \text{sign}\left(\sum_{\substack{i=\text{support} \\ \text{vectors}}} y_i \alpha_i G(x - x_i, 2\sigma^2) - b\right)$$

where α_i are the Lagrange multipliers, y_i are the class label, the kernel function is represented by the Gaussian distribution G , and b defines the hyperplane [135].

The strict constraint of searching for the perfect separation between classes can be softened by introducing a parameter C to tolerate errors.

SVM models were obtained with NeuroSolutions software [132].

To shorten the computational time required to select relevant descriptors, a pre-reduction step based on the model sensitivity was introduced. Sensitivity analysis is a method for extracting the cause-and-effect relationship between the inputs and outputs. After the learning phase the parameters are kept fixed. The basic idea is that the inputs are shifted slightly and the corresponding change in the output is the sensitivity of model outputs to the inputs. The SVM was trained with all inputs. This sensitivity analysis, which is different from sensitivity as a statistical

measure of classifiers, provides feedback about which inputs are the most significant. Those with a reduced or null effect on the outputs were removed to prune the number of initial variables. This process reduced the size of the input decreasing the complexity and the training time of the model.

After that, a GA procedure combined with SVM was performed by optimizing as fitness function the costs on the validation set. The specific parameters used for GA and SVM are reported in Table 5.3.

Table 5.3 Parameter setting used for GA and SVM.

GA parameters	
<u>GA setting</u>	Progression: generational (entire population is replaced with each iteration) Selection for the next generation: Roulette, rank-based crossover point number = 1, probability of crossover = 0.9 Mutation operator: uniform, probability of mutation = 0.01
<u>GA training options</u>	Nr. Epochs: 1000 Population size: 50 Max generation nr.: 100 Termination criteria: terminate after 100 epochs w/o improvement on validation set Class importance: use classes equally weighted
SVM parameters	
<u>Step</u>	Step size: GA optimisation in the range 0-1
<u>Supervised Learning Control</u>	Max nr. Epochs: 1000 Termination criteria: Increasing in MSE on the validation set Weight update: batch

5.2.3. PERFORMANCE EVALUATION

For all the methods the best model was selected on the basis of Copper statistics and the other parameters - already introduced in Chapter 4 (§4.1.2.1), for both training and validation sets. The robustness of the models was evaluated by cross-validation with a 10-fold leave-several-out (LSO). Then, the real prediction ability of the models was assessed with the help of the external test set never used to build or select the best model. To visualize better the models' behaviour the Receiver Operating Characteristic (ROC) curve was used to compare graphically

performances on the training, validation and test sets obtained with the different modelling techniques. ROC graph represents an alternative way to confusion matrices, to examine the classifier performances, by plotting 1-specificity versus sensitivity. ROC curves have proven to be a valuable way to evaluate the quality of a two-class classifier. The point (0,1) is the perfect classifier, as all positive and negative cases are predicted correctly. The points (0,0) and (1,1) represent a classifier that predicts all cases to be negative and positive, respectively, whereas (1,0) is associated with a classifier that always predicts wrongly. The closer is the model to the point (0, 1) the better it is. The main advantage in using the ROC graph is that it incorporates all information contained in the confusion matrix, since FN is the complement of TP and TN is the complement of FP. It also provides a visual tool for examining the trade-off between the ability of a classifier correctly to identify positive cases and the number of negative cases that are incorrectly classified. ROC curves for classifiers have been exemplified in the recent Predictive Toxicology Challenge [136].

5.3. RESULTS AND DISCUSSION

For each method the best model was selected on the basis of training and validation sets' performances and then verified on the chemicals in the test set. Particular attention was paid to the balance between sensitivity and specificity in the choice of the preferred model. The model characteristics for RBA are summarised in Table 5.4 and graphically compared through the ROC values in Figure 5.5. The selected descriptors (with the exclusion of those used in the DF model) are reported in Table 5.5. For all the methods it was possible to reach an accuracy around 85% or above, on all the three subsets. Due to the biased distribution in the two classes, sensitivity tends to be lower than specificity.

Table 5.4 Overview of RBA results.

	Training set (180P/326N)			Validation set (54P/96N)			Test set (54P/96N)		
	Acc.	Spec.	Sens.	Acc.	Spec.	Sens.	Acc.	Spec.	Sens.
DF	97.04	98.77	93.89	88.67	92.71	81.48	85.33	88.54	79.63
<i>77 descriptors in a combination of 3 trees</i>									
AFP	86.36	88.65	82.22	88.67	94.79	77.78	85.33	87.50	81.48
<i>6 descriptors (V64-V76-V103-V211-V221-V250) generating 24 rules</i>									
CART	85.38	84.36	87.22	84.00	87.50	77.78	85.33	85.42	85.19
<i>8 descriptors (V64-V97-V107-V119-V180-V182-V221-V250), 11 terminal nodes</i>									
MLP	84.39	89.26	75.56	87.33	91.67	79.63	84.00	87.50	77.78
<i>7 descriptors (V62-V70-V91-V111-V197-V221-V250)</i>									
SVM	89.92	96.32	78.33	87.33	90.63	81.48	86.67	93.75	74.07
<i>12 descriptors (V13-V23-V40-V59-V61-V77-V80-V83-V100-V103-V111-V221)</i>									

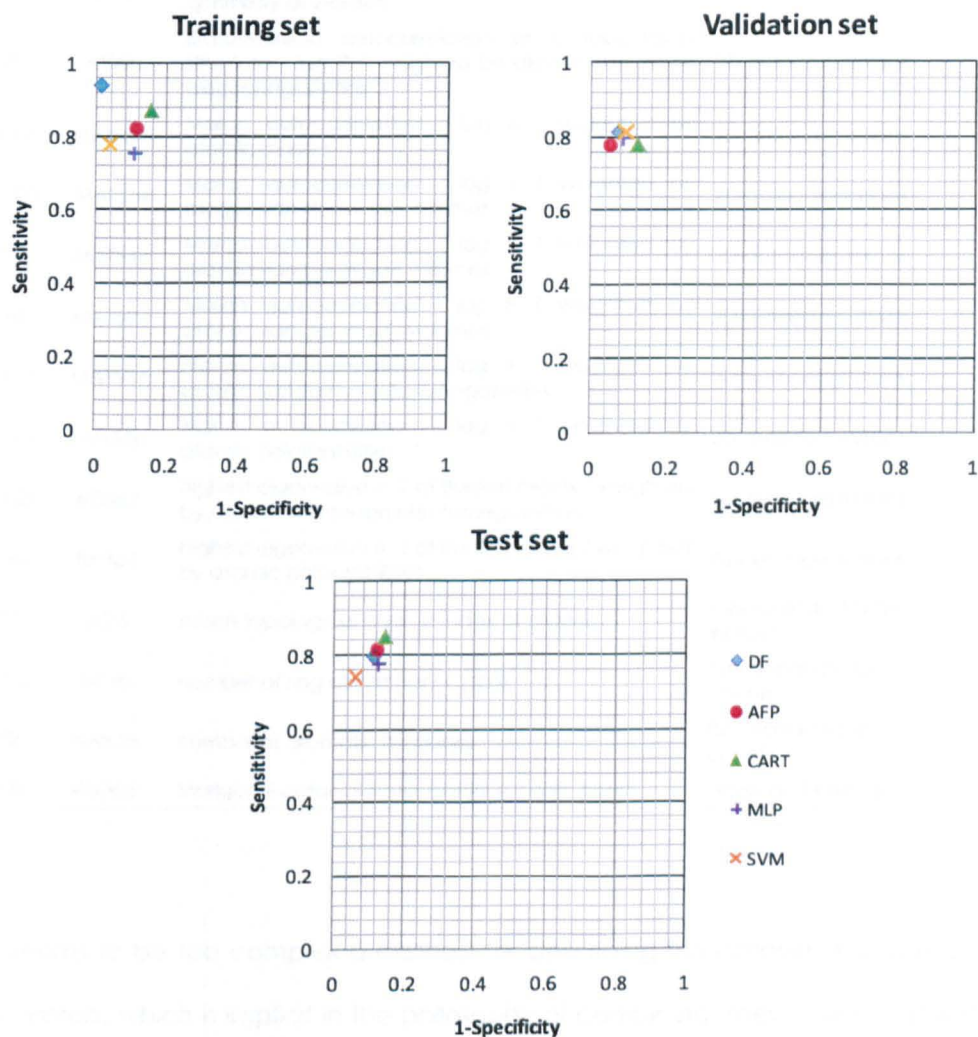
**Figure 5.5** ROC comparison of the RBA models obtained with different algorithms.

Table 5.5 List of selected variables in the RBA models.

No.	Symbol	Definition	Class
V13	NS	Number of Sulfur atoms	constitutional
V23	MSD	mean square distance index (Balaban)	topological
V40	PJ12	2D Pettitjean shape index	topological
V59	X0A	Average connectivity index chi-0	topological
V61	X2A	Average connectivity index chi-2	topological
V62	X4A	Average connectivity index chi-4	topological
V64	X2v	Valence connectivity Index chi-2	topological
V70	AAC	mean information index on atomic composition	information indices
V76	Vindex	Balaban V index	information indices
V77	SIC0	structural information content (neighborhood symmetry of 0-order)	information indices
V80	SIC1	structural information content (neighborhood symmetry of 1-order)	information indices
V83	SIC2	structural information content (neighborhood symmetry of 2-order)	information indices
V91	ATS8e	Broto-Moreau autocorrelation of a topological structure - lag 8 / weighted by atomic Sanderson electronegativities	2D autocorrelations
V97	MATS6m	Moran autocorrelation - lag 6 / weighted by atomic masses	2D autocorrelations
V100	MATS1v	Moran autocorrelation - lag 1 / weighted by atomic van der Waals volumes	2D autocorrelations
V103	MATS4v	Moran autocorrelation - lag 4 / weighted by atomic van der Waals volumes	2D autocorrelations
V107	MATS8v	Moran autocorrelation - lag 8 / weighted by atomic van der Waals volumes	2D autocorrelations
V111	MATS4e	Moran autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities	2D autocorrelations
V119	MATS5p	Moran autocorrelation - lag 5 / weighted by atomic polarizabilities	2D autocorrelations
V180	BEHe2	highest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities	Burden eigenvalues
V182	BEHp1	highest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities	Burden eigenvalues
V197	JGI6	mean topological charge index of order 6	topological charge indices
V211	NCr _q	number of ring quaternary C(sp ³)	functional group counts
V221	NArOH	number of aromatic hydroxyls	functional group counts
V250	MLOGP	Moriguchi octanol-water partition coeff. (logP)	physico-chemical

DF seems to be too complex a method for describing this dataset; it uses many descriptors, which is implicit in the philosophy of combining trees developed with diverse variables, and the performances in cross-validation are decisively lower.

This large gap between fitting (Acc. = 97%) and cross-validation (10-fold LSO Acc. = 74%) can be a symptom of overfitting. This hypothesis is in part confirmed by the drop in accuracy for test and validation sets whereas the other methods maintain a better stability of this parameter for these sets but also in cross-validation (e.g.: 10-fold LSO Acc. = 81% for AFP; 10-fold LSO Acc. = 75% for CART).

Above all, RBA models obtained with CART and AFP seem to be preferable achieving similar performances to the others but being based on a more intuitive syntax for the model codified in simple if-then rules.

The nArOH descriptor was selected in all models, while MLOGP was present in all but DF and SVM models. The different selection strategies seem to perform well, converging in the selection of relevant descriptors, already identified in the literature to describe oestrogenic effects. In fact, nArOH, the number of phenolic rings, has already been demonstrated to be a valuable descriptor since it accounts for the possibility to create H-bonds with the aminoacids and the water molecules in the binding pocket. Similarly, lipophilicity is considered important to describe the hydrophobic central region of the binding pocket. nArOH is the most relevant descriptor for CART – responsible for the first splitting, see Figure 5.6 – and it is also present in all the rules identified by AFP, whereas it is more difficult to ascribe the correct importance to the different descriptors for the other models, since the relationship they identify is not so explicit. DF descriptors are even more difficult to interpret in view of the multitude of variables used. Two other groups of descriptors are often selected: connectivity indices and 2D autocorrelation descriptors. Moran 2D autocorrelation descriptors are measures of spatial autocorrelation that can be weighted by different atomic properties; those more frequently selected in RBA models are related to bulky properties. Molecular connectivities are topological descriptors based on a count of groupings of skeletal atoms, weighted by degree of skeletal branching. Lower order indices are considered to encode mainly the bulk of a molecule, whereas higher order

indices encode more subtle features such as the presence of rings and branching patterns [137].

The CART model is preferable also in view of its simplicity and the tree is reported in Figure 5.6.

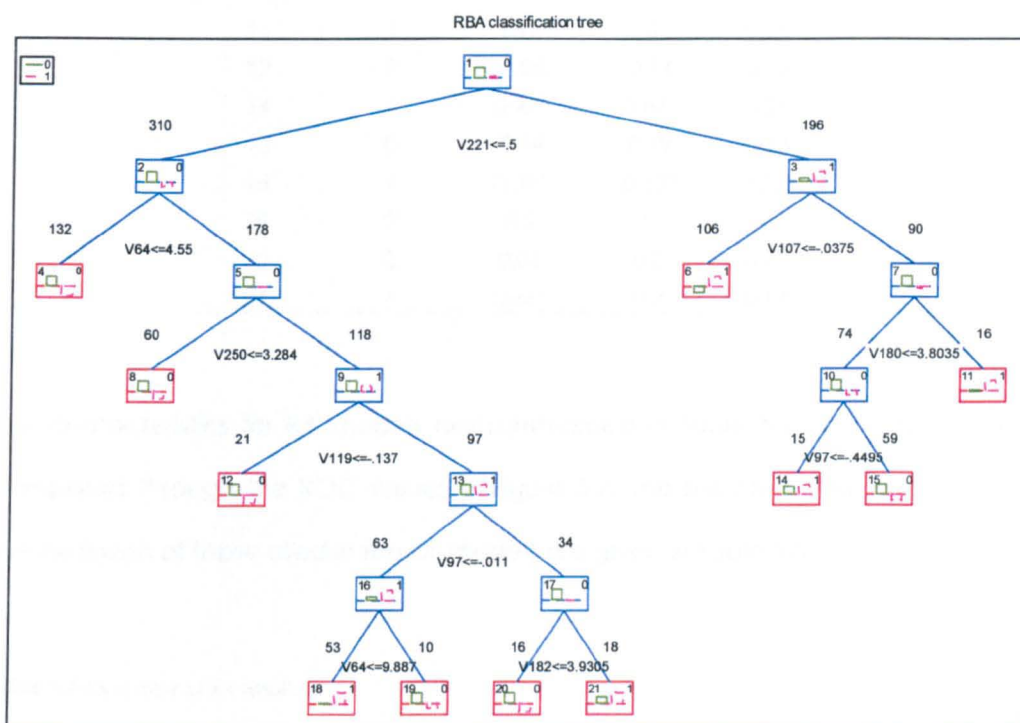


Figure 5.6 CART tree for RBA endpoint

In the CART tree the node assignment can be used to evaluate the reliability of the predictions: some of the terminal nodes are characterised by a lower purity grade and this behaviour is maintained also in the test set as shown in Table 5.6. By excluding the compound assigned through the two most impure nodes (14 and 18, both assigning compounds to the active class) the accuracy increases consistently in all the three sets, up to 90%. On the other hand, in this way specificity increases but sensitivity decreases.

Table 5.6 Misclassification rates for the CART terminal nodes for RBA. Node purity can be used to assess prediction reliability. Asterisks identify nodes with a misclassification rate greater than 0.3.

Terminal node	Node assignment	Misclassification ratio		
		Training set	Validation set	Test set
4	0	0.04	0.02	0
6	1	0.13	0.16	0.14
8	0	0.1	0.26	0.17
11	1	0.31*	0	0.25
12	0	0.05	0.14	0.25
14	1	0.47*	0.67*	0.25
15	0	0.14	0.19	0.13
18	1	0.32*	0.33*	0.55*
19	0	0.2	0	0
20	0	0.06	0.2	0.17
21	1	0.44*	0.2	0.17

The characteristics for RA models are summarised in Table 5.7 and graphically compared through the ROC values in Figure 5.7. The selected descriptors (with the exclusion of those used in the DF model) are given in Table 5.8.

Table 5.7 Overview of RA results.

	Training set (117P/389N)			Validation set (35P/115N)			Test set (35P/115N)		
	Acc.	Spec.	Sens.	Acc.	Spec.	Sens.	Acc.	Spec.	Sens.
DF	98.62	100.00	94.02	86.67	93.91	62.86	85.33	89.57	71.43
	<i>81 descriptors in a combination of 3 trees</i>								
AFP	87.35	92.29	70.94	89.33	92.17	80.00	88.67	92.17	77.14
	<i>6 descriptors (V74-V140-V189-V211-V221-V250) generating 29 rules</i>								
CART	86.76	87.15	85.47	81.33	86.96	62.86	83.33	88.70	65.71
	<i>9 descriptors (V11-V32-V35-V76-V162-V194-V221-V234-V250), 12 terminal nodes</i>								
MLP	87.94	91.26	76.92	95.33	97.39	88.57	86.67	91.30	71.43
	<i>8 descriptors (V4-V56-V59-V153-V182-V200-V221-V246)</i>								
SVM	98.62	100.00	94.02	91.33	97.39	71.43	80.67	93.04	40.00
	<i>10 descriptors (V2-V39-V70-V86-V130-V146-V188-V193-V195-V197)</i>								

For all the methods it was possible to reach an accuracy greater than 80%, on all the three subsets. Due to the low proportion of active compounds in the dataset

(25% only) the different methods have difficulty in maintaining a stable behaviour in the sensitivity trend. AFP and MLP seem less affected by this problem while SVM and, to a lower extent, also DF and CART have a larger drop of sensitivity in the validation and/or test sets.

Overall, RA models obtained with AFP and MLP seem to perform better. Again nArOH and MLOGP were often selected as relevant descriptors. Other descriptors present in all the models are those belonging to topological charge indices. They were proposed to evaluate the charge transfer between pairs of atoms and therefore the global charge transfer in the molecule.

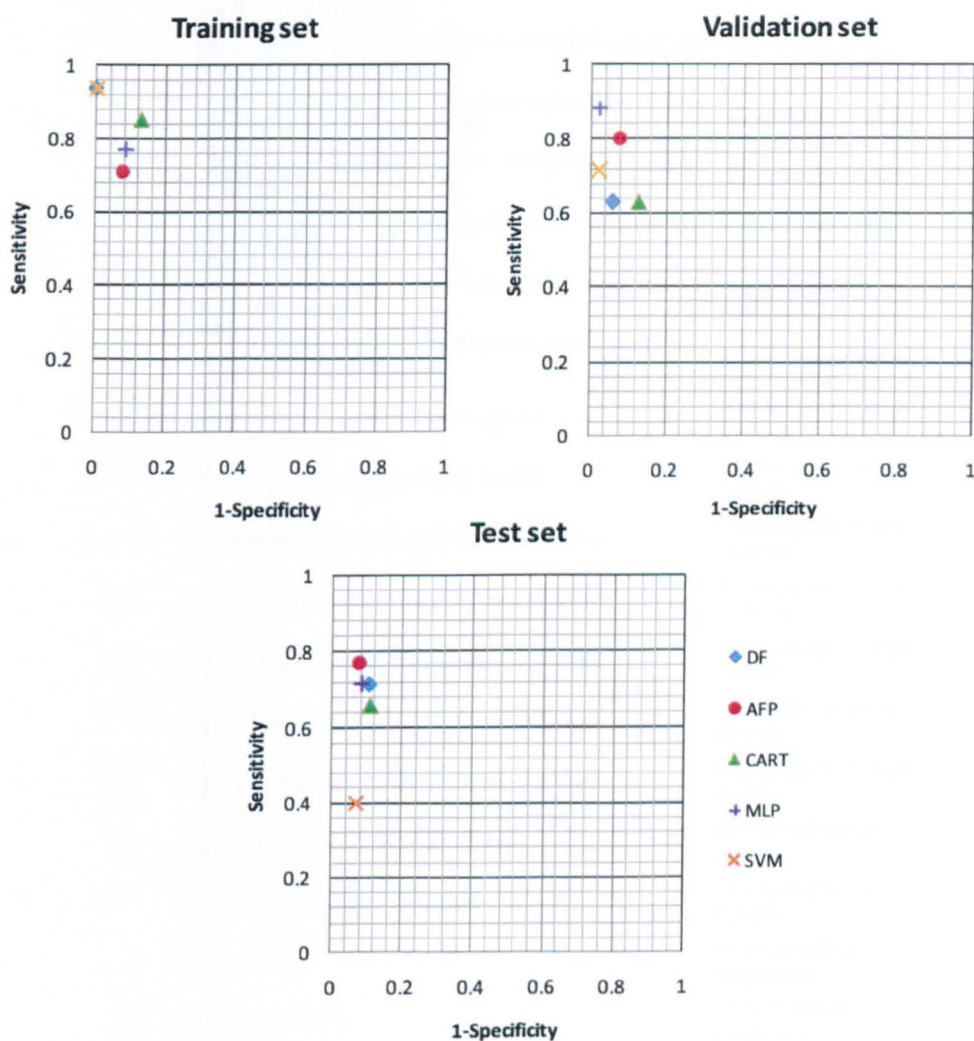


Figure 5.7 ROC comparison of the RA models obtained with different algorithms.

Table 5.8 List of selected variables in the RA models.

No.	Symbol	Definition	Class
V2	Mv	mean atomic van der Waals volume (scaled on Carbon atom)	constitutional
V4	Ms	mean electrotopological state	constitutional
V11	nN	number of Nitrogen atoms	constitutional
V32	MAXDN	maximal electrotopological negative variation	topological
V35	TIE	E-state topological parameter	topological
V39	PW5	path/walk 5 - Randic shape Index	topological
V56	plID	conventional bond-order ID number	walk and path counts
V59	X0A	average connectivity index chi-0	topological
V70	AAC	mean information index on atomic composition	information indices
V74	HVcpx	graph vertex complexity index	Information indices
V76	Vindex	Balaban V Index	information indices
V86	BIC3	bond information content (neighborhood symmetry of 3-order)	information indices
V130	GATS1v	Geary autocorrelation - lag 1 / weighted by atomic van der Waals volumes	2D autocorrelations
V140	GATS3e	Geary autocorrelation - lag 3 / weighted by atomic Sanderson electronegativities	2D autocorrelations
V146	GATS1p	Geary autocorrelation - lag 1 / weighted by atomic polarizabilities	2D autocorrelations
V153	EEig12x	Eigenvalue 12 from edge adj. matrix weighted by edge degrees	edge adjacency indices
V162	EEig14d	Eigenvalue 14 from edge adj. matrix weighted by dipole moments	edge adjacency indices
V182	BEHp1	highest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities	Burden eigenvalues
V188	GGI7	topological charge index of order 7	topological charge indices
V189	GGI8	topological charge index of order 8	topological charge indices
V193	JGI2	mean topological charge Index of order2	topological charge indices
V194	JGI3	mean topological charge index of order3	topological charge indices
V195	JGI4	mean topological charge index of order4	topological charge indices
V197	JGI6	mean topological charge index of order6	topological charge indices
V200	JGI9	mean topological charge index of order9	topological charge indices
V211	nCrq	number of ring quaternary C(sp3)	functional group counts
V221	nArOH	number of aromatic hydroxyls	functional group counts
V234	C-026	R-CX-R	atom-centred fragments
V246	N-075	R-N-R / R-N-X	atom-centred fragments
V250	MLOGP	Moriguchi octanol-water partition coeff. (logP)	physico-chemical

In the hypothesis of adopting binary classification as a screening tool to assess oestrogenic capabilities of chemical contaminants, emphasis should be placed to false negative results. In this context, lower sensitivity is not an advisable characteristic, so possible modification of this behaviour, exhibited by all classifiers for both endpoints, was also investigated, through combination of their prediction abilities. In the case of RBA a combination of CART and SVM was used. CART was chosen because it is characterised by the largest sensitivity on the training set (excluding the complex DF model), while SVM selected different FN from the other classifiers. Assigning a compound to the active class when at least one of the two models predicts it as active, increases the sensitivity significantly, maintaining an accuracy around 85% on all the three sets (see Table 5.9). Moreover the few FN compounds exhibit very marginal activity compared to those present in the single models, as highlighted in Figure 5.8. A similar situation can be proposed for RA: the SVM model is very specific, while MLP has a better sensitivity overall on the training and validation sets. The statistics improve even though sensitivity remains a little too poor on the test set (see Table 5.9). However, as already found for RBA, the experimental activities for FN present in the combined models are extremely low (one thousand times less than the reference compound) compared to those of FN in the single models (see Figure 5.9).

Table 5.9 Performances of the combined models for RBA and RA.

	Combined RBA model (CART + SVM)			Combined RA model (MLP + SVM)		
	Training set	Validation set	Test set	Training set	Validation set	Test set
FP	59	19	19	34	6	14
FN	8	3	3	3	2	9
TP	172	51	51	114	33	26
TN	267	77	77	355	109	101
Acc.	86.76	85.33	85.33	92.69	94.67	84.67
Spec.	81.90	80.21	80.21	91.26	94.78	87.83
Sens.	95.56	94.44	94.44	97.44	94.29	74.29

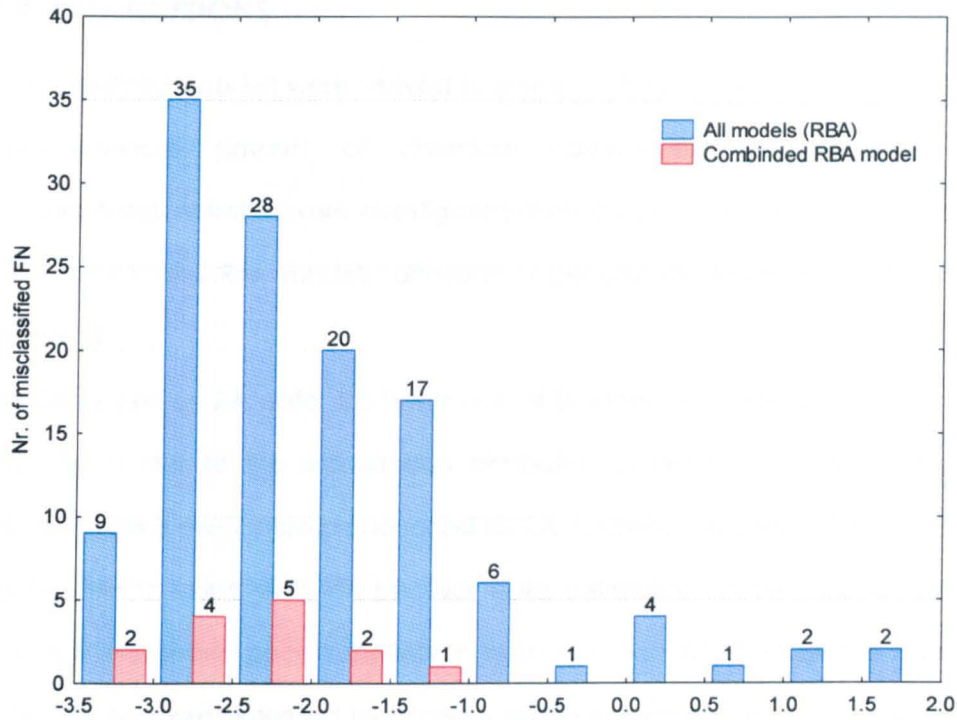


Figure 5.8 Comparison of misclassified FN for single and combined RBA models.

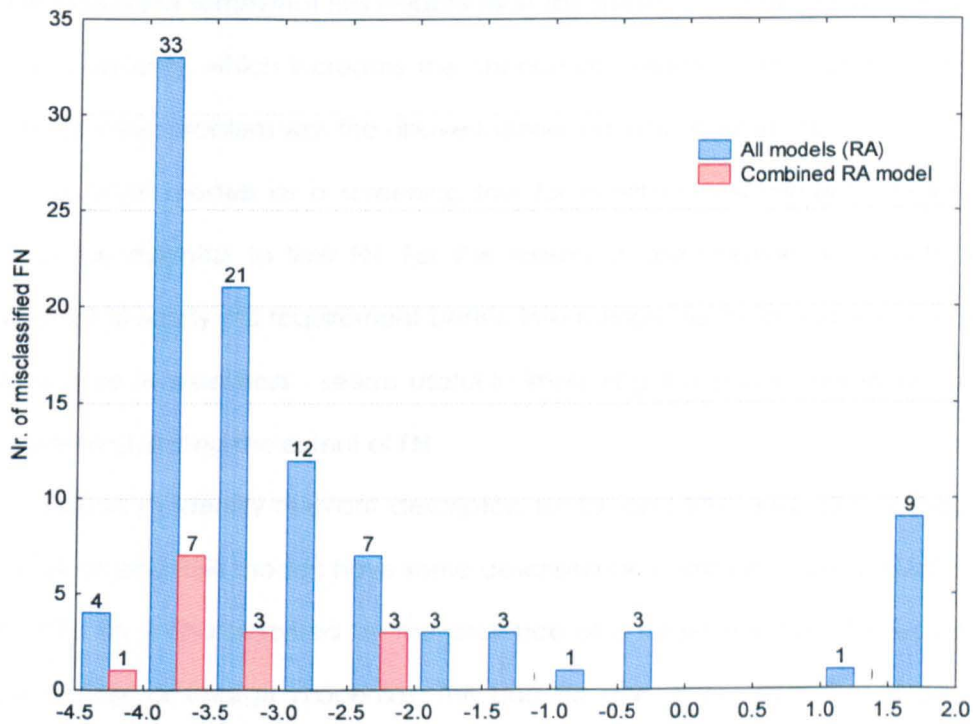


Figure 5.9 Comparison of misclassified FN for single and combined RA models.

5.4. CONCLUSIONS

New predictive models were derived to assess oestrogenicity for a very large and heterogeneous dataset of chemical compounds. Both binding and transcriptional activities were investigated and very good accuracy was reached for both RA and RBA models, although a general weakness in sensitivity was observed.

Performances on RA data are lower overall in terms of sensitivity than those on RBA. This is due to the unbalanced distribution of data in the two classes, a situation that is even more pronounced for RA. Furthermore, some of the methods seem able to cope with this peculiar data distribution better than others; for instance AFP shows good sensitivity in both RBA and RA. SVM on the contrary seems to be more disturbed by a poor class representation and consequently is characterised by lower sensitivity.

All the methods were capable of developing satisfactory models even though DF and SVM seem somewhat less reliable than the others. The main drawback of DF is its complexity, which increases the chance of overfitting, while in the case of SVM the main problem was the above-mentioned lack of sensitivity.

To use QSAR models as a screening tool for prioritising experimental testing, it would be essential to limit FN. For this reason a combination of models was explored to satisfy this requirement better. Interestingly, for both endpoints, SVM – apart from its weakness – seems useful in improving the performances of single models and limiting the extent of FN.

It is difficult to identify relevant descriptors for RA and RBA endpoints. Although overall RA and RBA models have some descriptors in common, such as NArOH or MLOGP, RA is characterised by the presence of a larger number of descriptors accounting for charge properties. This can be due to the fact that a proper interaction with the receptor is required to exhibit the transcriptional effect and

charge distribution is very important for that, once the ligand is in bound in the binding site.

This dataset is at the basis of MultiCASE models for receptor binding assay and reporter gene as reported in the Danish database [138]. Although not too many details are available on these commercial models, results in crossvalidation (LSO of 50% of data) reported there yield an accuracy around 80% with a sensitivity lower than 80% for both endpoints. The models here obtained are slightly better and especially the combined models are characterised by an increased sensitivity also on the external validation set.

Aside from the combined models, by globally evaluating both the performances and the easiness, the best RBA model is the CART classification tree; while, in the case of RA, the preferred model can be the one developed according to AFP algorithm.

Overall the models so obtained are sufficiently robust and characterised by ease of use for their application in the regulatory context.

CHAPTER 6

CONCLUSIONS

This work aimed to explore the status, availability and reliability of non-testing methods applied to endocrine disruption mediated by the oestrogen receptor, and eventually to propose new models easily exploitable in regulatory contexts.

Three existing QSAR models were selected in this work for a deeper evaluation on the basis of the OECD principles for QSAR validation. Different kinds of model (regression and classification, SAR and QSAR models) have been analysed in detail and they were externally validated with new data found in the literature. Models relying on bi-dimensional descriptors only seemed more user-friendly and more reproducible. The inclusion of 3D parameters – providing a more complete structural characterisation – required a detailed definition of the protocol used for their calculation and the evaluation of the prediction sensitivity to procedural steps. In the example presented in this study, the model was robust and reproducible to a satisfactory level. Several ways of assessing the applicability domain have been evaluated depending on the available information. Although some methods seem better than others, no efficient way to detect poor predictions has been identified.

This piece of work explored a quite unusual aspect of QSAR modelling: often new models appear in the literature addressing a similar or identical dataset already investigated by other researchers. Less frequently the focus is given to further applications of the model and the possibility to transfer it to other scientists is usually not considered. This does not necessarily mean that the model itself cannot be used practically by someone other than the original developer (all the

models examined here were quite satisfactory from this point of view), but neglecting this aspect may decrease the chance to use the model in the future due to a lack of reproducibility and accessibility.

Beside this necessary focus on what was already available in the scientific community, very useful indications were derived for continuation of the project by developing new classification models with ease of use as screening methods. In this framework, SAR or classification models can be a valid alternative to quantitative models. Further indications from this stage of the research were obtained: it appeared that to avoid overfitting especially when dealing with complex models, an accurate validation is essential, including an external test set. Moreover preference was given to the use of bi-dimensional descriptors in newly developed QSAR, since it was observed that similar performances can often be reached with the use of 2D and 3D descriptors respectively.

Bearing all these considerations in mind, data collected have been used to develop QSAR binary classifiers based on different data-mining techniques such as classification trees and decision forest, adaptive fuzzy partition (AFP), neural networks and support vector machines (SVM). New predictive models were derived to assess oestrogenicity for a very large and heterogeneous dataset of chemical compounds. Attention was focussed on multiple *in vitro* endpoints to characterize better the effects of EDs evaluating both binding (RBA) and transcriptional activity (RA).

A very good accuracy was achieved for both RA and RBA models (around 85% in all instances), although a general weakness in sensitivity was observed. Performances on RA data were lower overall in terms of sensitivity than those on RBA. This is due to the unbalanced distribution of data in the two classes, especially for RA. In addition, some of the methods seem able to cope with this peculiar data distribution better than others; for instance AFP shows good sensitivity in both RBA and RA. SVM on the other hand seems to be more

disturbed by a poor class representation and consequently is characterised by lower sensitivity.

In this context, some more complex model architecture has been explored, such as NN, for consideration as a benchmark for the best obtainable results.

Model performances were quite good, comparable to those available in the literature, and especially the combined models are slightly better, being characterised by an increased sensitivity. The validation procedure adopted was quite demanding, including both internal and external validation.

The descriptors were statistically selected and a convergence to certain specific descriptors was observed in this selection, which supports the theoretical explanation for the relevance of some of these descriptors - such as presence of a phenolic ring - in the underlying mechanism that controls the strength of oestrogenic effects.

Attention was focused on some specific characteristics, to achieve the objective of developing models for regulatory purposes. For this reason, a first aspect considered was that in using QSAR models as a screening tool for prioritising experimental testing it would be essential to limit FN. For this reason a combination of models was explored to satisfy this requirement better. Other aspects such as simplicity and reliability of the models, have been emphasised in view of possible application in the regulatory framework.

The errors observed were of a limited number and of limited extent and their presence is implicit in the underlying statistical approach at the basis of QSAR analysis. Nevertheless, any misclassified compound can be problematical for the acceptance of a model, and the definition of the applicability domain (AD) of a QSAR should help in addressing this issue. Some hints from the validation exercise indicate that similarity assessment is not sufficient to increase the reliability of predictions from QSAR models. The data here obtained during development of new models suggested that a more comprehensive way to define the AD would

be to include the model characteristics itself. The example of the CART model for RBA shows that some information about the reliability of predicted values can be derived by observing the specific node activated to generate the prediction. However, the AD concept requires further elucidation, particularly as regards to the applicative context. In the case of QSAR usage for regulatory purposes, still it has to be discussed if the accuracy reachable by QSAR would be considered acceptable by regulators.

Beside the aspect of AD, other issues can be further explored in future work. A possible extension of this work is to couple the classification approach adopted here with quantitative models so as to include in the analysis the magnitude of the activity. Another possible direction for further analysis is to restrict the chemical domain under investigation, so as to assess a more focused chemical space. This can be of help in deriving some more detailed mechanistic reasoning and it can be an advantage for characterizing the AD.

Overall the work conducted in this project holds out new predictive models addressing the effects of xenobiotics to the endocrine system, particularly mediated by the oestrogen receptor. They can be a valuable complement to *in vivo* and *in vitro* studies in the toxicological characterisation of chemical compounds.

The original parts of the project conducted by the PhD candidate were the following:

- Literature survey to search for promising models and data;
- Selection of the models on the basis of criteria derived from the OECD principles;
- Contribution to the datasets preparation for the validation and checking of chemical structures;

- Calculation of the majority of the descriptors needed to validate existing models;
- Elaboration of the results of the validation exercise;
- Data preparation and checking of structures for the dataset used for developing new models;
- Descriptor calculation;
- Dataset splitting in training test and validation set with self-organizing map;
- Development of binary classification models for oestrogenicity with classification trees, decision forest, AFP, neural network (MLP) and support vector machines;
- Discussion of the results and of the new models obtained.

REFERENCES

- [1] Danzo BJ. Environmental xenobiotics may disrupt normal endocrine function by interfering with the binding of physiological ligands to steroid receptors and binding proteins. *Environmental Health Perspectives* 1997;105 (3):294-301.
- [2] COM/99/0706 final. Communication from the Commission to the Council and the European Parliament - Community strategy for endocrine disrupters - A range of substances suspected of interfering with the hormone systems of humans and wildlife. 1999.
- [3] COM/2001/0262 final. Communication from the Commission to the Council and the European Parliament on the implementation of the Community Strategy for Endocrine Disrupters - a range of substances suspected of interfering with the hormone systems of humans and wildlife (COM (1999) 706). 2001.
- [4] European Commission. European workshop on the impact of endocrine disrupters on human health and wildlife: report of the proceedings. Weybridge, UK. EUR 17549. Weybridge, UK, 1997.
- [5] BKH Report. Towards the establishment of a priority list of substances for further evaluation of their role in endocrine disruption - preparation of a candidate list of substances as a basis for priority-setting. 2000.
- [6] Damstra T, Barlow S, Bergman A, Kavlock R, van der Kraak G, editors. Global assessment of the state-of-the-science of endocrine disruptors: ICPS-WHO, 2002.
- [7] Robinson-Rechavi M, Escriva Garcia H, Laudet V. The nuclear receptor superfamily. *Journal of Cell Science* 2003;116 (Part 4):585-6.
- [8] Giguere V. Orphan nuclear receptors: from gene to function. *Endocrine Reviews* 1999;20 (5):689-725.

-
- [9] Enmark E, Gustafsson JA. Oestrogen receptors - an overview. *Journal of Internal Medicine* 1999;246 (2):133-8.
- [10] Diel P. Tissue-specific estrogenic response and molecular mechanisms. *Toxicology Letters* 2002;127 (1-3):217-24.
- [11] Pearce ST, Jordan VC. The biological role of estrogen receptors [alpha] and [beta] in cancer. *Critical Reviews in Oncology/Hematology* 2004;50 (1):3-22.
- [12] Kuiper GG, Carlsson B, Grandien K, Enmark E, Haggblad J, Nilsson S, Gustafsson JA. Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. *Endocrinology* 1997;138 (3):863-70.
- [13] Nilsson S, Gustafsson JA. Estrogen receptor transcription and transactivation: Basic aspects of estrogen action. *Breast Cancer Research: BCR* 2000;2 (5):360-6.
- [14] McKenna NJ, O'Malley BW. An interactive course in nuclear receptor signaling: concepts and models. *Science Signaling STKE* 2005;2005 (299):tr22.
- [15] Nilsson S, Makela S, Treuter E, Tujague M, Thomsen J, Andersson G, Enmark E, Pettersson K, Warner M, Gustafsson et al. Mechanisms of estrogen action. *Physiological Reviews* 2001;81 (4):1535-65.
- [16] Mueller-Farnow A, Egner U. Ligand-binding domain of estrogen receptors. *Current Opinion in Biotechnology* 1999;10 (6):550-6.
- [17] Ruff M, Gangloff M, Wurtz JM, Moras D. Estrogen receptor transcription and transactivation: Structure-function relationship in DNA- and ligand-binding domains of estrogen receptors. *Breast Cancer Research: BCR* 2000;2 (5):353-9.
- [18] Klinge CM. Estrogen receptor interaction with co-activators and co-repressors. *Steroids* 2000;65 (5):227-51.

-
- [19] Brzozowski AM, Pike AC, Dauter Z, Hubbard RE, Bonn T, Engstrom O, Ohman L, Greene GL, Gustafsson JA, Carlquist et al. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* 1997;389 (6652):753-8.
- [20] Low LY, Hernandez H, Robinson CV, O'Brien R, Grossmann JG, Ladbury JE, Luisi B. Metal-dependent folding and stability of nuclear hormone receptor DNA-binding domains. *Journal of Molecular Biology* 2002;319 (1):87-106.
- [21] Nettles KW, Sun J, Radek JT, Sheng S, Rodriguez AL, Katzenellenbogen JA, Katzenellenbogen BS, Greene GL. Allosteric control of ligand selectivity between estrogen receptors [alpha] and [beta]: implications for other nuclear receptors. *Molecular Cell* 2004;13 (3):317-27.
- [22] EDSTAC. Endocrine disruptor screening and testing advisory committee final report. *Fed. Regist. Not.* 63 (248), 71541– 71568. 1998.
- [23] O'Connor JC, Cook JC, Marty MS, Davis LG, Kaplan AM, Carney EW. Evaluation of Tier I screening approaches for detecting endocrine-active compounds (EACs). *Critical Reviews in Toxicology* 2002;32 (6):521-49.
- [24] ICCVAM. Evaluation of in vitro test methods for detecting potential endocrine disruptors: estrogen receptor and androgen receptor binding and transcriptional activation assays. NIH Publication No. 03-4503. 2003.
- [25] Sanseau P. Impact of human genome sequencing for in silico target discovery. *Drug Discovery Today* 2001;6 (6):316-23.
- [26] Werner T, Nelson PJ. Joining high-throughput technology with in silico modelling advances genome-wide screening towards targeted discovery. *Briefings in Functional Genomics & Proteomics* 2006;5 (1):32-6.
- [27] Haley-Vicente D, Edwards DJ. Proteomic informatics: in silico methods lead to data management challenges. *Current Opinion in Drug Discovery & Development* 2003;6 (3):322-32.

-
- [28] Michalovich D, Overington J, Fagan R. Protein sequence analysis in silico: application of structure-based bioinformatics to genomic initiatives. *Current Opinion in Pharmacology* 2002;2 (5):574-80.
- [29] Mazzatorta P, Benfenati E, Lorenzini P, Vighi M. QSAR in ecotoxicity: an overview of modern classification techniques. *Journal of Chemical Information and Computer Sciences* 2004;44 (1):105-12.
- [30] Ren S. Ecotoxicity prediction using mechanism- and non-mechanism-based QSARs: a preliminary study. *Chemosphere* 2003;53 (9):1053-65.
- [31] Benigni R, Passerini L. Carcinogenicity of the aromatic amines: from structure-activity relationships to mechanisms of action and risk assessment. *Mutation Research* 2002;511 (3):191-206.
- [32] Livingstone DJ, Greenwood R, Rees R, Smith MD. Modelling mutagenicity using properties calculated by computational chemistry. *SAR and QSAR in Environmental Research* 2002;13 (1):21-33.
- [33] Patlewicz G, Rodford R, Walker JD. Quantitative structure-activity relationships for predicting mutagenicity and carcinogenicity. *Environmental Toxicology and Chemistry / SETAC* 2003;22 (8):1885-93.
- [34] Ivanova AA, Ivanov AA, Oliferenko AA, Palyulin VA, Zefirov NS. Highly diverse, massive organic data as explored by a composite QSPR strategy: an advanced study of boiling point. *SAR and QSAR in Environmental Research* 2005;16 (3):231-46.
- [35] Kahn I, Fara D, Karelson M, Maran U, Andersson PL. QSPR treatment of the soil sorption coefficients of organic pollutants. *Journal of Chemical Information and Modeling* 2005;45 (1):94-105.
- [36] Ekins S, Waller CL, Swaan PW, Cruciani G, Wrighton SA, Wikel JH. Progress in predicting human ADME parameters in silico. *Journal of Pharmacological and Toxicological Methods* 2000;44 (1):251-72.

-
- [37] Hansch C, Leo A, Meikapati SB, Kurup A. QSAR and ADME. *Bioorganic & Medicinal Chemistry* 2004;12 (12):3391-400.
- [38] Allanou R, Hansen BG, van der Bilt Y. Public availability of data on EU high production volume chemicals - Part 1. *Chemistry Today* 2003;21 (6):91-5.
- [39] Allanou R, Hansen BG, van der Bilt Y. Public availability of data on EU high production volume chemicals - Part 2. *Chemistry Today* 2003;21 (7/8):59-64.
- [40] EC 1907/2006. Regulation of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). 2006.
- [41] COM/2005/0007 final. Report from the Commission to the Council and the European Parliament - fourth report on the statistics on the number of animals used for experimental and other scientific purposes in the Member States of the European Union {SEC(2005) 45} 2005.
- [42] COM/2001/0088 final. White Paper - Strategy for a future chemicals policy 2001.
- [43] Pedersen F, de Bruijn J, Munn S, van Leeuwen K. Assessment of additional testing needs under REACH. Effects of (Q)SARS, risk based testing and voluntary industry initiatives. EUR 20863 EN. 2003.
- [44] van der Jagt K, Munn S, Tørsløv J, de Bruijn J. Alternative approaches can reduce the use of test animals under REACH. Addendum to the report: Assessment of additional testing needs under REACH Effects of (Q)SARS, risk based testing and voluntary industry initiatives. EUR 21405 EN. 2004.
- [45] Worth AP, Hartung T, Van Leeuwen CJ. The role of the European centre for the validation of alternative methods (ECVAM) in the validation of (Q)SARs. *SAR and QSAR in Environmental Research* 2004;15 (5-6):345-58.
- [46] Cronin MT, Jaworska JS, Walker JD, Comber MH, Watts C.D., A.P. W. Use of QSARs in international decision-making frameworks to predict health

- effects of chemical substances. *Environmental Health Perspectives* 2003;111 (10):1391-401.
- [47] Cronin MT, Walker JD, Jaworska JS, Comber MH, Watts C.D., A.P. W. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. *Environmental Health Perspectives* 2003;111 (10):1376-90
- [48] Fang H, Tong W, Shi LM, Blair R, Perkins R, Branham W, Hass BS, Xie Q, Dial SL, Moland et al. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chemical Research in Toxicology* 2001;14 (3):280-94.
- [49] Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, Branham WS, Dial SL, Moland CL, Sheehan et al. QSAR models using a large diverse set of estrogens. *Journal of Chemical Information and Computer Sciences* 2001;41 (1):186-95.
- [50] Yu SJ, Keenan SM, Tong W, Welsh WJ. Influence of the structural diversity of data sets on the statistical quality of three-dimensional quantitative structure-activity relationship (3D-QSAR) models: predicting the estrogenic activity of xenoestrogens. *Chemical Research in Toxicology* 2002;15 (10):1229-34.
- [51] Hong H, Tong W, Fang H, Shi L, Xie Q, Wu J, Perkins R, Walker JD, Branham W, Sheehan et al. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environmental Health Perspectives* 2002;110 (1):29-36.
- [52] Tong W, Perkins R, Fang H, Hong H, Xie Q, Branham W, Sheehan DM, Anson JF. Development of quantitative structure-activity relationships (QSARs) and their use for priority setting in the testing strategy of endocrine disruptors. *Regulatory Research Perspectives* 2002;1 (3):1-16.

-
- [53] Waller CL, Oprea TI, Chae K, Park HK, Korach KS, Laws SC, Wiese TE, Kelce WR, Gray LE, et al. Ligand-based identification of environmental estrogens. *Chemical Research in Toxicology* 1996;9 (8):1240-8.
- [54] Hansch C. Quantitative approach to biochemical structure-activity relationships. *Accounts of Chemical Research* 1969;2 (8):232-9.
- [55] Kubinyi H. From narcosis to hyperspace: the history of QSAR. *Quantitative Structure-Activity Relationships* 2002;21 (4):348-56.
- [56] McFarland JW. Parabolic relation between drug potency and hydrophobicity. *Journal of Medicinal Chemistry* 1970;13 (6):1192-6.
- [57] Todeschini R, Consonni V. *Handbook of molecular descriptors*. Weinheim: WILEY-VCH, 2000.
- [58] Leach AR. *Molecular modelling: principles and applications*. Upper Saddle River, NJ: Prentice Hall, 2001.
- [59] Benfenati E, Piclin N, Roncaglioni A, Vari MR. Factors influencing predictive models for toxicology. *SAR and QSAR in Environmental Research* 2001;12 (6):593-603.
- [60] Bradbury SP, Mekenyan OG, Ankley GT. The role of ligand flexibility in predicting biological activity: structure-activity relationships for aryl hydrocarbon, estrogen and androgen receptor binding affinity. *Environmental Toxicology and Chemistry* 1998;17 (1):15-25.
- [61] Ghafourian T, Cronin MTD. The impact of variable selection on the modelling of oestrogenicity. *SAR and QSAR in Environmental Research* 2005;16 (1-2):171-90.
- [62] Piclin N, Pintore M, Wechman C, Roncaglioni A, Benfenati E, Chretien JR. Ecotoxicity prediction by adaptive fuzzy partitioning: comparing descriptors computed on 2D and 3D structures. *SAR and QSAR in Environmental Research* 2006;17 (2):225-51.

-
- [63] Devillers J, editor. Neural networks in QSAR and drug design. London, UK: Academic Press, 1996.
- [64] Golbraikh A, Tropsha A. Beware of q²! Journal of Molecular Graphics & Modelling 2002;20 (4):269-76.
- [65] Tropsha A, Gramatica P, Gombar Vijay K. The importance of being earnest; validation is the absolute essential for successful application and interpretation of QSPR models. QSAR & Combinatorial Science 2003;22 (1):69-77.
- [66] Hawkins DM. The problem of overfitting. Journal of Chemical Information and Modeling 2004;44 (1):1-12.
- [67] Golbraikh A, Shen M, Xiao Z, Xiao Y-D, Lee K-H, Tropsha A. Rational selection of training and test sets for the development of validated QSAR models. Journal of Computer-Aided Molecular Design 2003;17 (2-4):241-53.
- [68] Golbraikh A, Tropsha A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. Journal of Computer-Aided Molecular Design 2002;16 (5-6):357-69.
- [69] Netzeva TI, Worth A, Aldenberg T, Benigni R, Cronin MTD, Gramatica P, Jaworska JS, Kahn S, Klopman G, Marchant et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. Alternatives To Laboratory Animals: ATLA 2005;33 (2):155-73.
- [70] Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. Alternatives To Laboratory Animals: ATLA 2005;33 (5):445-59.
- [71] Klopman G, Chakravarti SK. Structure-activity relationship study of a diverse set of estrogen receptor ligands (I) using MultiCASE expert system. Chemosphere 2003;51 (6):445-59.

-
- [72] Gao H, Williams C, Labute P, Bajorath J. Binary quantitative structure-activity relationship (QSAR) analysis of estrogen receptor ligands. *Journal of Chemical Information and Computer Sciences* 1999;39 (1):164-8.
- [73] Tong W, Xie Q, Hong H, Shi L, Fang H, Perkins R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environmental Health Perspectives* 2004;112 (12):1249-54.
- [74] Sutherland JJ, O'Brien LA, Weaver DF. Spline-fitting with a genetic algorithm: a method for developing classification structure-activity relationships. *Journal of Chemical Information and Computer Sciences* 2003;43 (6):1906-15.
- [75] Marini F, Roncaglioni A, Novic M. Variable selection and interpretation in structure-affinity correlation modeling of estrogen receptor binders. *Journal of Chemical Information and Modeling* 2005;45 (6):1507-19.
- [76] Asikainen AH, Ruuskanen J, Tuppurainen KA. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR and QSAR in Environmental Research* 2004;15 (1):19-32.
- [77] Zheng W, Tropsha A. Novel variable selection quantitative structure-property relationship approach based on the k-Nearest-Neighbor principle. *Journal of Chemical Information and Modeling* 2000;40 (1):185-94.
- [78] Kaiser Klaus LE. Neural networks for effect prediction in environmental and health issues using large datasets. *QSAR & Combinatorial Science* 2003;22 (2):185-90.
- [79] Schmieder PK, Ankley G, Mekenyan O, Walker JD, Bradbury S. Quantitative structure-activity relationship models for prediction of

- estrogen receptor binding affinity of structurally diverse chemicals. *Environmental Toxicology and Chemistry / SETAC* 2003;22 (8):1844-54.
- [80] Netzeva TI, Gallegos Saliner A, Worth AP. Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. *Environmental Toxicology and Chemistry / SETAC* 2006;25 (5):1223-30.
- [81] Cunningham AR, Cunningham SL, Rosenkranz HS. Structure-activity approach to the identification of environmental estrogens: the MCASE approach. *SAR and QSAR in Environmental Research* 2004;15 (1):55-67.
- [82] Gallegos Saliner A, Amat L, Carbo-Dorca R, Schultz TW, Cronin MTD. Molecular quantum similarity analysis of estrogenic activity. *Journal of Chemical Information and Computer Sciences* 2003;43 (4):1166-76.
- [83] Brown N, McKay B, Gasteiger J. Fingal: a novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modelling. *QSAR & Combinatorial Science* 2005;24 (4):480-4.
- [84] Gao H, Lajiness MS, Van Drie J. Enhancement of binary QSAR analysis by a GA-based variable selection method. *Journal of Molecular Graphics & Modelling* 2002;20 (4):259-68.
- [85] Cruciani G, editor. *Molecular interaction fields: applications in drug discovery and ADME prediction*. Weinheim: Wiley-VCH, 2006.
- [86] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society* 1988;110 (18):5959-67.
- [87] Akamatsu M. Current state and perspectives of 3D-QSAR. *Current Topics in Medicinal Chemistry* 2002;2 (12):1381-94.
- [88] Martin YC. 3D QSAR: current state, scope, and limitations. *Perspectives in Drug Discovery and Design* 1998;V12-14 (0):3-23.

-
- [89] Waller CL. A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *Journal of Chemical Information and Modeling* 2004;44 (2):758-65.
- [90] Tong W, Perkins R, Xing L, Welsh WJ, Sheehan DM. QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. *Endocrinology* 1997;138 (9):4022-5.
- [91] Coleman Kelly P, Toscano William A, Wiese JThomas E. QSAR models of the in vitro estrogen activity of bisphenol A analogs. *QSAR & Combinatorial Science* 2003;22 (1):78-88.
- [92] Koshland-Jr DE. The joys and vicissitudes of protein science. *Protein Science* 1993;2 (8):1364-8.
- [93] Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. *Journal of Computer-Aided Molecular Design* 2002;V16 (3):151-66.
- [94] Oprea TI, Marshall GR. Receptor-based prediction of binding affinities. *Perspectives in Drug Discovery and Design* 1998;V9-11 (0):35-61.
- [95] Shoichet BK. Virtual screening of chemical libraries. *Nature* 2004;432 (7019):862-5.
- [96] Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *Journal of Medicinal Chemistry* 2006;49 (20):5851-5.
- [97] Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* 2006;49 (20):5912-31.
- [98] vanLipzig MMH, terLaak AM, Jongejan A, Vermeulen NPE, Wamelink M, Geerke D, Meerman JHN. Prediction of ligand binding affinity and

- orientation of xenoestrogens to the estrogen receptor by molecular dynamics simulations and the linear interaction energy method. *Journal of Medicinal Chemistry* 2004;47 (4):1018-30.
- [99] DeLisle RK, Yu S-J, Nair AC, Welsh WJ. Homology modeling of the estrogen receptor subtype [beta] (ER-[beta]) and calculation of ligand binding affinities. *Journal of Molecular Graphics and Modelling* 2001;20 (2):155-67.
- [100] Yoon S, Welsh WJ. Identification of a minimal subset of receptor conformations for improved multiple conformation docking and two-step scoring. *Journal of Chemical Information and Modeling* 2004;44 (1):88-96.
- [101] Wade RC, Henrich S, Wang T. Using 3D protein structures to derive 3D-QSARs. *Drug Discovery Today: Technologies* 2004;1 (3):241-6.
- [102] Wolohan P, Reichert DE. CoMFA and docking study of novel estrogen receptor subtype selective ligands. *Journal of Computer-Aided Molecular Design* 2003;V17 (5):313-28.
- [103] Sippl W. Receptor-based 3D QSAR analysis of estrogen receptor ligands – merging the accuracy of receptor-based alignments with the computational efficiency of ligand-based methods. *Journal of Computer-Aided Molecular Design* 2000;V14 (6):559-72.
- [104] Vedani A, Dobler M, Lill MA. Combining protein modeling and 6D-QSAR. simulating the binding of structurally diverse ligands to the estrogen receptor. *Journal of Medicinal Chemistry* 2005;48 (11):3700-3.
- [105] Comber MHI, Walker JD, Watts C, Hermens J. Quantitative structure-activity relationships for predicting potential ecological hazard of organic chemicals for use in regulatory risk assessments. *Environmental Toxicology and Chemistry* 2003;22 (8):1822-8.
- [106] Gerner I, Spielmann H, Hoefler T, Liebsch M, Herzler M. Regulatory use of (Q)SARs in toxicological hazard assessment strategies. SAR and QSAR in *Environmental Research* 2004;15 (5):359 - 66.

-
- [107] Tunkel J, Mayo K, Austin C, Hickerson A, Howard P. Practical considerations on the use of predictive models for regulatory purposes. *Environmental Science & Technology* 2005;39 (7):2188-99.
- [108] Hartung T, Bremer S, Casati S, Coecke S, Corvi R, Fortaner S, Gribaldo L, Halder M, Hoffmann S, Roi et al. A modular approach to the ECVAM principles on test validity. *Alternatives To Laboratory Animals: ATLA* 2004;32 (5):467-72.
- [109] Witorsch RJ. Endocrine disruptors: can biological effects and environmental risks be predicted? *Regulatory Toxicology and Pharmacology* 2002;36 (1):118-30.
- [110] Tong W, Fang H, Hong H, Xie Q, Perkins R, Sheehan D. Receptor-mediated toxicity: QSARs for oestrogen receptor binding and priority setting of potential oestrogenic endocrine disruptors. In: Cronin MTD, Livingstone D, editors. *Predicting chemical toxicity and fate*. Boca Raton, FL, 2005. pp. 285-314.
- [111] Tong W, Fang H, Williams CR, Burch JM, Richard AM. DSSTox National Center for Toxicological Research Estrogen Receptor Binding Database (NCTRER): SDF files and website documentation. Updated version: NCTRER_v2a_232_1Mar05, www.epa.gov/ncct/dsstox/. 2003.
- [112] TSAR version 3.3. San Diego, CA: Accelrys Inc., 2000.
- [113] QSARis, version 1.1 San Diego, CA: SciVision - Academic Press, 2005.
- [114] Statistica, version 6.1. Vigonza, Italy: StatSoft Italia Srl, 2003.
- [115] Cooper JA, Saracci R, P. C. Describing the validity of carcinogen screening tests. *British Journal of Cancer* 1979;39:87-9.
- [116] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta* 1975;405:442-51.

-
- [117] Kaur H, Raghava GP. A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* 2004;20:2751-8.
- [118] CODESSA, verion 2.21. Gainesville, FL: University of Florida, 1995.
- [119] Mazzatorta P, Benfenati E, Schuller B, Romberg M, McCourt D, Dubitzky W, Sild S, Karelson M, Papp A, Bágyi I, Darvas F. OpenMolGRID: molecular science and engineering in a grid context. *Proceedings of PDPTA 2004, the 2004 international conference on parallel and distributed processing techniques and applications*. Las Vegas, NV, USA, 2004.
- [120] METI, Ministry of Economy Trade and Industry, Japan. Current status of testing methods development for endocrine disrupters. 6th Meeting of the task force on endocrine disrupters testing and assessment (EDTA), 24-25 June 2002, Tokyo, Japan. 2002.
- [121] Dragon, version 5.4. Milan, Italy: Talete Srl, 2005.
- [122] Nikolova-Jeliazkova N, Jaworska J. AmbitDisclosure, a QSAR applicability domain assessment software. <http://ambit.acad.bg/>.
- [123] Stewart JJP. MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design* 1990;4:1-103.
- [124] Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. Belmont, CA: Wadsworth, 1984.
- [125] Ripley BD. *Pattern recognition and neural networks*. Cambridge: Cambridge University Press, 1996.
- [126] Tong W, Hong H, Fang H, Xie Q, Perkins R. Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences* 2003;43 (2):525-31.
- [127] Ros F, Pintore M, Chretien JR. Molecular descriptor selection combining genetic algorithms and fuzzy logic: application to database mining

- procedures. *Chemometrics and Intelligent Laboratory Systems* 2002;63 (1):15-26.
- [128] Ros F, Taboureau O, Pintore M, Chretien JR. Development of predictive models by adaptive fuzzy partitioning. Application to compounds active on the central nervous system. *Chemometrics and Intelligent Laboratory Systems* 2003;67 (1):29-50.
- [129] Pintore M, Taboureau O, Ros F, Chretien JR. Database mining applied to central nervous system (CNS) activity. *European Journal of Medicinal Chemistry* 2001;36 (4):349-59.
- [130] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 1958;65 (6):386-408.
- [131] Haykin S. *Neural networks: a comprehensive foundation* (2nd Edition). Englewood Cliffs: Prentice-Hall, Inc., 1998.
- [132] NeuroSolutions, version 5.06. Gainesville, FL: NeuroDimension, Inc., 2005.
- [133] Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 1998;2 (2):121-67.
- [134] Johansson P, Ringnér M. Classification of genomic and proteomic data using support vector machines. In: Dubitzky W, Granzow M, Berrar DP, editors. *Fundamentals of data mining in genomics and proteomics*. Berlin: Springer, 2007.
- [135] Bennett K, P, Campbell C. Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.* 2000;2 (2):1-13.
- [136] Toivonen H, Srinivasan A, King RD, Kramer S, Helma C. Statistical evaluation of the predictive toxicology challenge 2000-2001. *Bioinformatics* 2003;19 (10):1183-93.
- [137] Cronin MTD, Schultz TW. Pitfalls in QSAR. *Journal of Molecular Structure: THEOCHEM* 2003;622 (1-2):39-51.

- [138] Danish (Q)SAR Database, user manual for the internet version of the Danish (Q)SAR database, Version 1 May 2005, http://130.226.165.14/User_Manual_Danish_Database.pdf.

LIST OF ABBREVIATIONS

2D	Bi-dimensional
3D	Three-dimensional
3D-QSAR	Three-dimensional QSAR
AD	Applicability Domain
ADMET	Adsorption, Distribution, Metabolism, Excretion and Toxicity
AF	Activation Function
AFP	Adaptive Fuzzy Partition
CART	Classification and Regression Tree
CNS	Central Nervous System
CoMFA	Comparative Molecular Field Analysis
CoMSIA	Comparative Molecular Similarity Indices Analysis
DBD	DNA binding domain
DF	Decision Forest
E2	17 β -estradiol
ECVAM	European Centre for the Validation of Alternative Methods
ED	Endocrine Disrupter
EDKB	Endocrine Disruptor Knowledge Base
ER	Oestrogen Receptor
ERE	Estrogen Response Elements
EU	European Union
FC	Fuzzy Clustering
FL	Fuzzy Logic
FN	False Negatives

FP	False Positives
GA	Genetic Algorithm
HPV	High Production Volume substances, exceeding a production volume of 1000 t/year in Europe
HSA	Hybrid Selection Algorithm
HTS	High-Throughput Screening
ICCVAM	Interagency Coordinating Committee on the Validation of Alternative Methods
LBD	Ligand Binding Domain
LSO	Leave Several out
MCC	Matthews Correlation Coefficient
METI	Ministry of Economy, Trade and Industry (Japan)
MLP	Multi Layer Perceptron
MLR	Multi Linear Regression
MQSM	Molecular Quantum Similarity Measures
NCTRER	National Center for Toxicological Research - Estrogen Receptor Binding Database
NN	Neural Network
NR	Nuclear Receptor
NTP	National Toxicology Program
OECD	Organisation for Economic Co-operation and Development
PAH	Polycyclic Aromatic Hydrocarbon
PCA	Principal Component Analysis
PCB	Polychlorinated Biphenyls
PLS	Partial Least Squares
POP	Persistent Organic Pollutants
QSAR	Quantitative Structure-Activity Relationships

RA	Relative Activity
RBA	Relative Binding Affinity
RBF	Radial Basis Function
REACH	Registration, Evaluation, Authorisation and restriction of CHemicals
RMSE	Root Mean Squared Error
SAR	Structure-Activity Relationships
SIMCA	Soft Independent Modelling by Class Analogy
SVM	Support Vector Machines
TN	True negatives
TP	True Positives

ANNEX A

This annex contains the NTP dataset provided as supplementary material by Sutherland *et al.* [74] used in Chapter 4 for the validation of *Model 2*.

Features assigned according to the SAR model are reported. Compounds with light gray background were those belonging to the training set. Compounds with activity class U (undefined) were those excluded from this evaluation.

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Acetamide; Anilide	tox-469	Metolachlor	51218-45-2		1	1	0	0	0	0	
Acrylate	tox-386	2-Hydroxyethyl methacrylate	868-77-9	0	0	-	-	-	-	-	0
Acrylate	tox-438	Methyl methacrylate	80-62-6	0	0	-	-	-	-	-	0
Acrylate	tox-621	Triethylene glycol dimethacrylate	109-16-0	0	0	-	-	-	-	-	0
Acrylate	tox-72	Bisphenol A diglycidyl ether dimethacrylate	1565-94-2	0	1	1	0	-	-	-	0
Acrylate	tox-75	Bisphenol A ethoxylate diacrylate	64401-02-1	U	-	-	-	-	-	-	-
Acrylate; Bisphenol	tox-73	Bisphenol A dimethacrylate	3253-39-2	1	1	1	0	-	-	0	0
Alcohol	tox-498	1,8-Octanediol	629-41-4		0	0	0	0	0	0	
Alcohol	tox-55	Benzyl alcohol	100-51-6		1	1	0	0	0	0	
Alcohol	tox-364	n -Hexanol	111-27-3		0	0	0	0	0	0	
Alcohol	tox-433	MER-25	67-98-1	U	-	-	-	-	-	-	-
Aldehyde	tox-347	Heptanal	111-71-7		0	0	0	0	0	0	
Alkoxyphenol	tox-428	Isoeugenol	97-54-1		1	1	1	0	0	0	
Alkoxyphenol	tox-631	Vanillin	121-33-5		1	1	1	0	0	0	
Alkoxyphenol	tox-314	Eugenol	97-53-0		1	1	1	0	0	0	
Alkoxyphenol	tox-346	4-(Heptyloxy)phenol	13037-86-0		1	1	1	0	0	1	
Alkoxyphenol	tox-367	Hexestrol monomethyl ether	13026-26-1		1	1	1	0	1	0	
Alkylbenzene	tox-92	sec - Butylbenzene	135-98-8		1	1	0	0	0	0	
Alkylphenol	tox-499	4-n - Octylphenol	1806-26-4		1	1	1	0	0	0	
Alkylphenol	tox-500	4-tert - Octylphenol	140-66-9		1	1	1	0	0	0	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Alkylphenol	tox-259	Dopamine	51-61-6		1	1	1	0	0	0	
Alkylphenol	tox-22	4-tert - Amylphenol	80-46-6		1	1	1	0	0	1	
Alkylphenol	tox-525	4- Phenethylphenol	6335-83-7		1	1	1	0	1	0	
Alkylphenol	tox-308	2-Ethylphenol	90-00-6		1	1	1	0	0	0	
Alkylphenol	tox-309	3-Ethylphenol	620-17-7		1	1	1	0	0	1	
Alkylphenol	tox-257	4- Dodecylphenol	104-43-8		1	1	1	0	0	1	
Alkylphenol	tox-175	2,6-Di-tert - butylphenol	128-39-2	0	1	1	1	-	-	0	0
Amide	tox-133	Colchicine	64-86-8	0	1	1	0	-	-	0	0
Anilide	tox-4	Alachlor	15972-60-8		1	1	0	0	0	0	
Aniline	tox-455	4,4'- Methylenebis(N,N- dimethylaniline)	101-61-1		1	1	0	0	1	0	
Aniline	tox-90	Butyl 4- aminobenzoate	94-25-7		1	1	0	0	0	0	
Aniline	tox-454	4,4'- Methylene dianiline	101-77-9		1	1	0	0	0	0	
Aniline	tox-21	4- Aminophenyl ether	101-80-4		1	1	0	0	0	0	
Aromatic heterocycle	tox-432	Melatonin	73-31-4		1	1	0	0	0	0	
Aromatic hydrocarbon	tox-489	Nonylbenzene	1081-77-2	0	1	1	0	-	-	0	0
Aromatic hydrocarbon; Alkylbenzene	tox-91	n - Butylbenzene	104-51-8		1	1	0	0	0	0	
Azo compound	tox-16	Amaranth	915-67-3		1	1	1	0	1	0	
Benzophenone	tox-208	2,2'- Dihydroxy-4- methoxybenzophenone	131-53-3		1	1	1	0	1	0	
Benzophenone	tox-209	2,2'- Dihydroxybenzophenone	835-11-0		1	1	1	0	1	0	
Benzophenone	tox-394	2-Hydroxy-4- methoxybenzophenone	131-57-7		1	1	1	0	1	0	
Benzophenone	tox-210	2,4- Dihydroxybenzophenone	131-56-6		1	1	1	0	1	0	
Benzophenone	tox-607	2,2',4,4'- Tetrahydroxybenzil	5394-98-9		1	1	1	0	1	0	
Benzophenone; Phenol	tox-214	4,4'- Dihydroxybenzophenone	611-99-4		1	1	1	0	1	0	
Biphenyl; Phenol	tox-549	2- Phenylphenol	90-43-7		1	1	1	0	0	0	
Biphenyl; Phenol	tox-551	4- Phenylphenol	92-69-3		1	1	1	0	0	1	
Biphenyl; Phenol	tox-550	3- Phenylphenol	580-51-8		1	1	1	0	0	0	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Biphenyldiol	tox-215	4,4'-Dihydroxybiphenyl	92-88-6	0	1	1	1	-	0	1	1
Bisphenol	tox-69	Bisphenol A	80-05-7		1	1	1	0	1	0	
Bisphenol	tox-85	Bisphenol S	80-09-1		1	1	1	0	1	0	
Bisphenol	tox-493	Nordihydroguariaric acid	500-38-9		1	1	1	0	1	0	
Bisphenol	tox-83	2,2'-Bisphenol F	2467-02-9		1	1	1	0	1	0	
Bisphenol	tox-138	Cyclofenil diphenol	5189-40-2	1	1	1	1	-	1	-	1
Bisphenol	tox-271	Erythro -MEA	20576-52-7	1	1	1	1	-	1	-	1
Bisphenol	tox-529	Phenol, 4,4'-[1,2-bis(methylene)-1,2-ethanediyl]bis-	107144-81-0	1	1	1	1	-	1	-	1
Bisphenol	tox-67	2,2-Bis(4-hydroxyphenyl)propanol	142648-65-5	U	-	-	-	-	-	-	-
Bisphenol	tox-70	Bisphenol A bis(chloroformate)	2024-88-6	1	1	1	0	-	-	0	0
Bisphenol	tox-71	Bisphenol A diglycidyl ether	1675-54-3	0	1	1	0	-	-	0	0
Bisphenol	tox-77	Bisphenol A propoxylate	37353-75-6	0	1	1	0	-	-	0	0
Bisphenol; Glucuronide	tox-76	Bisphenol A glucuronide		0	1	1	0	-	-	0	0
Bisphenol; Stilbene	tox-562	Pseudodiethyl stilbestrol	39011-86-4	1	1	1	1	-	1	-	1
Carbamate	tox-103	Carbofuran	1563-66-2		1	1	0	0	0	0	
Carbamate; Imidazole	tox-40	Benomyl	17804-35-2	0	1	1	0	-	-	0	0
Carbamate; Polycyclic aromatic hydrocarbon	tox-102	Carbaryl	63-25-2		1	1	0	0	0	0	
Carboxylic acid	tox-576	Suberic acid	505-48-6		0	0	0	0	0	0	
Carboxylic acid	tox-129	Cinnamic acid	621-82-9		1	1	0	0	0	0	
Carboxylic acid	tox-306	3-Ethyl-4-(p-methoxyphenyl)-2-methyl-3-cyclohexene-1-carboxylic acid	1755-52-8	1	1	1	0	-	-	0	0
Chalconoid	tox-552	Phloretin	60-82-2		1	1	1	0	1	0	
Chalconoid	tox-371	4-Hydroxychalcone	20426-12-4		1	1	1	0	1	0	
Chalconoid	tox-372	4'-Hydroxychalcone	2657-25-2		1	1	1	0	1	0	
Chalconoid	tox-622	4,2',4'-Trihydroxychalcone	961-29-5		1	1	1	0	1	0	
Chalconoid	tox-106	Chalcone	94-41-7		1	1	0	0	0	1	
Chalconoid	tox-373	4'-	38239-52-0	U	-	-	-	-	-	-	-

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		Hydroxychalcone (cis- and trans-)									
Chlorinated aromatic hydrocarbon; Organochlorine	tox-619	2,4,5-Trichlorophenoxyacetic acid	93-76-5		1	1	0	0	0	0	
Chlorinated cycloalkane; Organochlorine	tox-431	Lindane	58-89-9		1	0	0	0	0	0	
Chlorinated phenol	tox-122	2-Chlorophenol	95-57-8		1	1	1	0	0	0	
Chlorinated phenol	tox-120	2-Chloro-4-methylphenol	6640-27-3		1	1	1	0	0	1	
Chlorinated phenol	tox-123	4-Chlorophenol	106-48-9	U	-	-	-	-	-	-	-
Coumarin; Phenol	tox-136	Coumestrol	479-13-0		1	1	1	0	1	0	
Crown ether	tox-168	Dibenzo-18-crown-6	14187-32-7		1	1	0	0	0	0	
Cyclodiene	tox-6	Aldrin	309-00-2		1	0	0	0	0	0	
Dioxin	tox-603	2,3,7,8-Tetrachlorodibenzo-p-dioxin	1746-01-6	0	1	1	0	-	-	0	0
Diphenolalkane	tox-38	Aurin	603-45-2		1	1	1	0	1	0	
Diphenolalkane	tox-248	Diphenolic acid	126-00-1		1	1	1	0	1	0	
Diphenolalkane	tox-213	3,3'-Dihydroxyhexestrol	79199-51-2		1	1	1	0	1	0	
Diphenolalkane	tox-239	2,6-Dimethylhexestrol	334707-28-7		1	1	1	0	1	0	
Diphenolalkane	tox-366	DL-Hexestrol	5776-72-7	1	1	1	1	-	1	-	1
Diphenolalkane	tox-62	1,1-Bis-(4-hydroxyphenyl)ethane	2081-08-5	U	-	-	-	-	-	-	-
Diphenolalkane	tox-63	4,4-Bis(4-hydroxyphenyl)heptane	7425-79-8	1	1	1	1	-	1	-	1
Diphenolalkane	tox-64	3,4-Bis(3-hydroxyphenyl)hexane	68266-24-0	1	1	1	1	-	1	-	1
Diphenolalkane; Bisphenol	tox-365	Hexestrol	84-16-2		1	1	1	0	1	0	
Diphenolalkane; Bisphenol	tox-541	Phenol Red	143-74-8		1	1	1	0	1	0	
Diphenolalkane; Bisphenol	tox-79	Bisphenol B	77-40-7		1	1	1	0	1	0	
Diphenolalkane; Bisphenol	tox-82	Bisphenol E	6052-84-2		1	1	1	0	1	0	
Diphenolalkane; Bisphenol	tox-81	Bisphenol C 2	14868-03-2		1	1	1	0	1	0	
Diphenolalkane; Bisphenol	tox-42	Benzeneacetonitrile α-[bis(4-hydroxyphenyl)methylene]	66422-14-8	1	1	1	1	-	1	-	1
Diphenolalkane; Bisphenol	tox-65	3,3-Bis(4-hydroxyphenyl)pentane	3600-64-4	1	1	1	1	-	1	-	1
Diphenolalkane;	tox-66	1,1-Bis(4-	1576-13-2	1	1	1	1	-	1	-	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Bisphenol		hydroxyphenylpropane									
Diphenolalkane; Bisphenol	tox-78	Bisphenol AF	1478-61-1	1	1	1	1	-	1	-	1
Diphenolalkane; Bisphenol	tox-80	Bisphenol C	79-97-0	1	1	1	1	-	1	-	1
Diphenylalkene	tox-195	Dienestrol	84-17-3		1	1	1	0	1	0	
Diphenylalkene	tox-249	trans,trans - 1,4-Diphenyl-1,3-butadiene	886-65-7		1	1	0	0	0	0	
Diphenylalkene	tox-197	b-Dienestrol	35495-11-5	1	1	1	1	-	1	-	1
Ester	tox-199	Di-2-ethylhexyl adipate	103-23-1		0	0	0	0	0	0	
Flavanone	tox-320	Flavanone	17002-31-2		1	1	0	0	0	0	
Flavanone	tox-349	Hesperetin	520-33-2		1	1	1	0	1	0	
Flavanone	tox-105	(±)-Catechin	7295-85-4		1	1	1	0	1	0	
Flavanone	tox-387	3'-Hydroxyflavanone	92496-65-6		1	1	1	0	1	0	
Flavanone	tox-479	Naringenin	480-41-1		1	1	1	0	1	0	
Flavanone	tox-389	6-Hydroxyflavanone	4250-77-5		1	1	1	0	1	0	
Flavanone	tox-390	7-Hydroxyflavanone	6515-36-2		1	1	1	0	1	0	
Flavanone	tox-579	Taxifolin	480-18-2		1	1	1	0	1	0	
Flavanone	tox-480	Naringin	10236-47-2		1	1	1	0	1	0	
Flavanone	tox-388	4'-Hydroxyflavanone	135413-27-3		1	1	1	0	1	0	
Flavone	tox-565	Quercetin	117-39-5		1	1	1	0	1	0	
Flavone	tox-319	Fisetin	528-48-3		1	1	1	0	1	0	
Flavone	tox-321	Flavone	525-82-6		1	1	0	0	0	0	
Flavone	tox-571	Rutin	153-18-4		1	1	1	0	1	0	
Flavone	tox-391	6-Hydroxyflavone	6665-83-4		1	1	1	0	1	0	
Flavone	tox-392	7-Hydroxyflavone	6665-86-7		1	1	1	0	1	0	
Flavone	tox-395	6-Hydroxy-2'-methoxyflavone	61546-59-6		1	1	1	0	1	0	
Flavone	tox-39	Baicalein	491-67-8		1	1	1	0	1	0	
Flavone	tox-34	Apigenin	520-36-5		1	1	1	0	1	0	
Flavone	tox-127	Chrysin	480-40-0		1	1	1	0	1	0	
Flavone	tox-429	Kaempferol	520-18-3		1	1	1	0	1	0	
Flavone	tox-475	Morin	480-16-0		1	1	1	0	1	0	
Flavone	tox-477	Myricetin	529-44-2		1	1	1	0	1	0	
Flavone	tox-220	6,4'-Dihydroxyflavone	63046-09-3		1	1	1	0	1	0	
Flavone	tox-623	3,6,4'-Trihydroxyflavone	253195-19-6		1	1	1	0	1	0	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Heterocyclic aromatic aldehyde	tox-327	Furfural	98-01-1		1	1	0	0	0	0	
Indane	tox-404	Indanestrol	71855-45-3	1	1	1	1	-	1	-	1
Indene	tox-201	1,3-Diethyl-4-hydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-202	1,3-Diethyl-6-hydroxy 2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-251	2,3-Diphenylindene-1		1	1	1	0	-	-	0	0
Indene	tox-295	3-Ethyl-6,4'-dihydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-302	3-Ethyl-4'-hydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-303	3-Ethyl-6-hydroxy 2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-304	3-Ethyl-4'-hydroxy 2-phenylindene-1		1	1	1	1	-	1	-	1
Indene	tox-305	3-Ethyl-6-hydroxy 2-phenylindene-1		1	1	1	1	-	1	-	1
Indene	tox-316	2-(2-Fluorophenyl)-3-phenyl-6-hydroxyindene		1	1	1	1	-	1	-	1
Indene	tox-543	2-Phenyl-3-(2-fluoro-4-hydroxyphenyl)-6-hydroxyindene		1	1	1	1	-	1	-	1
Indene	tox-544	2-Phenyl-3-(2-fluorophenyl)-6-hydroxyindene		1	1	1	1	-	1	-	1
Indene	tox-545	3-Phenyl-4'-hydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-546	3-Phenyl-6-hydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Indene	tox-547	2-Phenyl-3-(2-methylphenyl)-6-hydroxyindene		1	1	1	1	-	1	-	1
Indene	tox-548	2-Phenyl-3-(4-methylphenyl)-6-hydroxyindene		1	1	1	1	-	1	-	1
Isoflavone	tox-328	Genistein	446-72-0		1	1	1	0	1	0	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Hel	F5 P3P	F6 Other	Predicted Class
Isoflavone	tox-270	Equol	531-95-3		1	1	1	0	1	0	
Isoflavone	tox-142	Daidzein	486-66-8		1	1	1	0	1	0	
Isoflavone	tox-326	Formononetin	485-72-3		1	1	1	0	1	0	
Isoflavone	tox-329	Genistin	529-59-9		1	1	1	0	1	0	
Isoflavone	tox-58	Biochanin A	491-80-5		1	1	1	0	1	0	
Isoflavone	tox-625	7,3',4'-Trihydroxyisoflavone	485-63-2		1	1	1	0	1	0	
Isoflavone	tox-624	6,7,4'-Trihydroxyisoflavone	17817-31-1		1	1	1	0	1	0	
Isoflavone	tox-561	Prunetin	552-59-0		1	1	1	0	1	0	
Isoflavone	tox-226	Dihydrogenistein	21554-71-2	1	1	1	1	-	1	-	1
Isoflavone	tox-330	Glyceollin	66241-09-6	1	1	1	1	-	1	-	1
Isoflavone	tox-331	Glycitein	40957-83-3	1	1	1	1	-	1	-	1
Isoflavone	tox-332	Glycitin	40246-10-4	U	-	-	-	-	-	-	-
Isoflavone	tox-427	Ipriflavone	35212-22-7	0	1	1	0	-	-	0	0
Nitrobenzene	tox-393	Hydroxyflutamide	52806-53-8	U	-	-	-	-	-	-	-
Nitrogen heterocycle	tox-426	Indole[3,2-b]carbazole		U	-	-	-	-	-	-	-
Organochlorine	tox-442	p,p' - Methoxychlor	72-43-5		1	1	0	0	1	0	
Organochlorine	tox-107	Chlordane	57-74-9		1	0	0	0	0	0	
Organochlorine	tox-470	Mirex	2385-85-5		1	0	0	0	0	0	
Organochlorine	tox-350	Hexachlorobenzene	118-74-1		1	1	1	0	0	0	
Organochlorine	tox-430	Kepone	143-50-0		1	0	0	1	0	1	
Organochlorine	tox-632	Vinclozolin	50471-44-8		1	1	0	0	0	0	
Organochlorine	tox-472	Monohydroxy methoxychlor	28463-03-8		1	1	1	0	1	0	
Organochlorine	tox-473	Monohydroxy methoxychlor olefin	75938-34-0		1	1	1	0	1	0	
Organochlorine	tox-436	Methoxychlor olefin	2132-70-9		1	1	0	0	0	0	
Organochlorine	tox-108	α-Chlordane	5103-71-9	0	1	0	-	0	-	-	0
Organochlorine	tox-109	Chlormequat chloride	999-81-5	0	0	-	-	-	-	-	0
Organochlorine	tox-143	m,p' -DDD	4329-12-8	0	1	1	0	-	-	0	0
Organochlorine	tox-191	3,5-Dichloro 2-hydroxy-2-methylbut-3-enalide	16776-82-1	U	-	-	-	-	-	-	-
Organochlorine	tox-193	2-[[[(3,5-Dichlorophenyl)amino]-carbamoyl]oxyl]-2-methyl-3-butenic acid	119209-27-7	U	-	-	-	-	-	-	-
Organochlorine	tox-262	α-Endosulfan	959-98-8	0	1	0	-	1	-	0	0
Organochlorine	tox-441	o,p' - Methoxychlor	30667-99-3	1	1	1	0	-	-	0	0
Organochlorine	tox-612	Toxaphene	8001-35-2	U	-	-	-	-	-	-	-
Organochlorine	tox-629	Tris(4-chlorophenyl)	27575-78-6	1	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		methane									
Organochlorine	tox-630	Tris(4-chlorophenyl) methanol	30100-80-8	1	1	1	0	-	-	1	1
Organochlorine; Bisphenol	tox-68	2,2-Bis(p-hydroxyphenyl)-1,1,1-trichloroethane	2971-36-0		1	1	1	0	1	0	
Organochlorine; Chlorinated cyclodiene	tox-333	Heptachlor	76-44-8		1	0	0	0	0	0	
Organochlorine; Chlorinated cyclodiene	tox-263	a,b-Endosulfan	115-29-7		1	0	0	1	0	0	
Organochlorine; Chlorinated cyclodiene	tox-194	Dieldrin	60-57-1		1	0	0	1	0	0	
Organochlorine; Chlorinated cyclodiene	tox-264	b-Endosulfan	33213-65-9	U	-	-	-	-	-	-	-
Organochlorine; Chlorinated phenol	tox-121	4-Chloro-2-methylphenol	1570-64-5		1	1	1	0	0	1	
Organochlorine; Diphenyl ether	tox-315	Fenvalerate	51630-58-1	0	1	1	0	-	-	0	0
Organochlorine; Diphenylalkane	tox-170	o,p'-DDE	3424-82-6		1	1	0	0	0	1	
Organochlorine; Diphenylalkane	tox-144	o,p'-DDD	53-19-0		1	1	0	0	0	1	
Organochlorine; Diphenylalkane	tox-171	o,p'-DDT	789-02-6		1	1	0	0	0	1	
Organochlorine; Diphenylalkane	tox-172	p,p'-DDD	72-54-8		1	1	0	0	1	0	
Organochlorine; Diphenylalkane	tox-173	p,p'-DDE	72-55-9		1	1	0	0	1	0	
Organochlorine; Diphenylalkane	tox-174	p,p'-DDT	50-29-3		1	1	0	0	1	0	
Organochlorine; Nitrile; Diphenyl ether	tox-141	Cypermethrin	52315-07-8	0	1	1	0	-	-	0	0
Organochlorine; Phenol	tox-113	4-Chloro-4'-biphenylol	28034-99-3		1	1	1	0	0	0	
Organochlorine; Phenol	tox-114	4-Chloro-m-cresol	59-50-7		1	1	1	0	0	1	
Organochlorine; Phenol	tox-111	2'-Chloro-4,4'-biphenyldiol	56858-70-9	1	1	1	1	-	0	1	1
Organochlorine; Phenol	tox-112	2-Chloro-4-biphenylol	23719-22-4	1	1	1	1	-	0	1	1
Organochlorine; Triazine	tox-572	Simazine	122-34-9		1	1	0	0	0	0	
Organochlorine; Chlorinated bridged cycloalkene	tox-487	cis-Nonachlor	5103-73-1	0	1	0	-	0	-	-	0
Organochlorine; Chlorinated bridged cycloalkene	tox-488	trans-Nonachlor	39765-80-5	0	1	0	-	0	-	-	0
Paraben	tox-94	Butylparaben	94-26-8		1	1	1	0	0	1	
Paraben	tox-348	Heptyl 4-paraben	1085-12-7		1	1	1	0	0	1	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Paraben	tox-296	2-Ethylhexyl paraben	5153-25-3		1	1	1	0	0	1	
Paraben	tox-57	Benzylparaben	94-18-8		1	1	1	0	1	0	
Paraben	tox-559	Propyl paraben	94-13-3		1	1	1	0	0	1	
Paraben	tox-307	Ethyl paraben	120-47-8		1	1	1	0	0	1	
Paraben	tox-439	Methyl paraben	99-76-3		1	1	1	0	0	1	
Phenol	tox-95	2-sec - Butylphenol	89-72-5		1	1	1	0	0	0	
Phenol	tox-98	4-sec - Butylphenol	99-71-8		1	1	1	0	0	0	
Phenol	tox-99	4-tert - Butylphenol	98-54-4		1	1	1	0	0	0	
Phenol	tox-203	meso-p -(a,b-Diethyl-p -methylphenethyl)phenol	267408-76-4		1	1	1	0	1	0	
Phenol	tox-56	4-Benzyloxyphenol	103-16-2		1	1	1	0	1	0	
Phenol	tox-137	p -Cumyl phenol	599-64-4		1	1	1	0	1	0	
Phenol	tox-1	4,4'-(1,3-Adamantane diyl)diphenol		1	1	1	1	-	0	1	1
Phenol	tox-2	2-(1-Adamantyl)-4-methylphenol	41031-50-9	0	1	1	1	-	0	1	1
Phenol	tox-3	4-(1-Adamantyl)phenol	29799-07-3	1	1	1	1	-	0	1	1
Phenol	tox-96	2-tert - Butylphenol	88-18-6	U	-	-	-	-	-	-	-
Phenol	tox-97	3-tert - Butylphenol	585-34-2	U	-	-	-	-	-	-	-
Phenol; Alkylphenol	tox-491	n - Nonylphenol	25154-52-3		1	1	1	0	0	1	
Phenol; Alkylphenol	tox-310	4-Ethylphenol	123-07-9		1	1	1	0	0	1	
Phenol; Alkylphenol	tox-490	p - Nonylphenol	104-40-5	1	1	1	1	-	0	1	1
Phenol; Alkylphenol	tox-611	Tosyl nonylphenol (mixed branched isomers)		U	-	-	-	-	-	-	-
Phenol; Bisphenol	tox-84	4,4'-Bisphenol F	620-92-8		1	1	1	0	1	0	
Phenoxy carboxylic acid	tox-192	2,4-Dichlorophenoxyacetic acid	94-75-7		1	1	0	0	0	0	
Phenyl ether	tox-87	1,3-Butanediol, 4-[4-(1,2,3,4-tetrahydro-6-hydroxy-2-phenyl-1-naphthalenyl)phenoxy]	107144-85-4	1	1	1	1	-	1	-	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Phenyl ether	tox-88	1,3-Butanediol, 4-[4-{1,2,3,4-tetrahydro-6-methoxy-2-phenyl-1-naphthyl}p-henoxy]	107163-56-4	1	1	1	0	-	-	0	0
Phenylalkene	tox-294	Ethyl cinnamate	103-36-6		1	1	0	0	0	0	
Phosphate ester	tox-628	Triphenyl phosphate	115-86-6		1	1	0	0	0	0	
Phthalate	tox-200	Diethylhexyl phthalate	117-81-7		1	1	0	0	0	0	
Phthalate	tox-204	Diethyl phthalate	84-66-2		1	1	0	0	0	0	
Phthalate	tox-93	Butyl benzyl phthalate	85-68-7		1	1	0	0	0	1	
Phthalate	tox-176	Dibutyl phthalate	84-74-2		1	1	1	0	0	0	
Phthalate	tox-247	Di-n-octyl phthalate	117-84-0		1	1	0	0	0	0	
Phthalate	tox-241	Dimethyl phthalate	131-11-3		1	1	0	0	0	0	
Phthalate	tox-232	Diisononyl phthalate	28553-12-0		1	1	0	0	0	0	
Phthalate	tox-229	Diisobutyl phthalate	84-69-5		1	1	0	0	0	0	
Phthalate	tox-100	Butyl phthalyl n-butyl glycolate	85-70-1	0	1	1	0	-	-	0	0
Phthalate	tox-169	Dibutyl benzyl phthalate		0	1	1	0	-	-	0	0
Phthalate	tox-225	Diethyl phthalate	84-75-3	0	1	1	0	-	-	0	0
Phthalate	tox-230	Diisodecyl phthalate	26761-40-0	0	1	1	0	-	-	0	0
Phthalate	tox-231	Diisoheptyl phthalate	41451-28-9	0	1	1	0	-	-	0	0
Phthalimide	tox-609	Thalidomide	50-35-1		1	1	0	0	0	0	
Polychlorinated biphenyl	tox-601	2',3',4',5'-Tetrachloro-4-biphenylol	67651-34-7		1	1	1	0	0	0	
Polychlorinated biphenyl	tox-184	2',5'-Dichloro-4-biphenylol	53905-28-5		1	1	1	0	0	0	
Polychlorinated biphenyl	tox-597	3,3',5,5'-Tetrachloro-4,4'-biphenyldiol	13049-13-3		1	1	1	0	0	0	
Polychlorinated biphenyl	tox-177	2,4'-Dichlorobiphenyl	34883-43-7		1	1	0	0	0	1	
Polychlorinated biphenyl	tox-584	2,2',4,4'-Tetrachlorobiphenyl	2437-79-8		1	1	0	0	0	0	
Polychlorinated biphenyl	tox-594	3,3',4,4'-Tetrachlorobiphenyl	32598-13-3		1	1	0	0	0	0	
Polychlorinated biphenyl	tox-181	4,4'-Dichlorobiphenyl	2050-68-2		1	1	0	0	0	0	
Polychlorinated biphenyl	tox-178	2,5-Dichlorobiphenyl	34883-39-1	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		nyl									
Polychlorinated biphenyl	tox-179	3,4-Dichlorobiphenyl	2974-92-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-180	3,5-Dichlorobiphenyl	34883-41-5	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-182	2,5-Dichloro-2'-biphenylol	53905-30-9	0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-183	2,5-Dichloro-3'-biphenylol	53905-29-6	U	-	-	-	-	-	-	-
Polychlorinated biphenyl	tox-185	2,6-Dichloro-4'-biphenylol	79881-33-7	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-186	3,4-Dichloro-2'-biphenylol	209613-97-8	0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-187	3,4-Dichloro-3'-biphenylol	14962-34-6	0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-188	3,4-Dichloro-4'-biphenylol	53890-77-0	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-189	3,5-Dichloro-2'-biphenylol		0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-190	3,5-Dichloro-4'-biphenylol		0	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-334	2,2',3,3',4',5,5'-Heptachloro-4-biphenylol	158076-64-3	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-335	2,2',3,3',4,5,6-Heptachlorobiphenyl	68194-16-1	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-336	2,2',3,3',4',5,6-Heptachlorobiphenyl	52663-70-4	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-337	2,2',3,3',4,5,6-Heptachlorobiphenyl	52663-64-6	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-338	2,2',3',4,4',5,5'-Heptachloro-3-biphenylol	158076-69-8	1	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-339	2,2',3,4,4',5',6-Heptachlorobiphenyl	52663-69-1	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-340	2,2',3,4,4',6,6'-Heptachlorobiphenyl	74472-48-3	1	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-341	2,2',3,4',5,5',6-Heptachloro-4-biphenylol	158076-68-7	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-342	2,2',3,4',5,5',6-Heptachlorobiphenyl	52663-68-0	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-343	2,2',3,4',5,6,6'-Heptachlorobiphenyl	74487-85-7	1	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-344	2,3,3',4,4',5,6-Heptachlorobiphenyl	41411-64-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-345	2,3,3',4',5,5',6-Heptachlorobiphenyl	69782-91-8	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-351	2,2',3,3',4,4'-Hexachlorobiphenyl	38380-07-3	0	1	1	0	-	-	0	0
Polychlorinated	tox-352	2,2',3,4,4',5'-	35065-28-2	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
biphenyl		Hexachlorobi phenyl									
Polychlorinated biphenyl	tox-353	2,2',3,4,5,6'-Hexachlorobi phenyl	68194-15-0	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-354	2,2',3,4',5',6'-Hexachlorobi phenyl	38380-04-0	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-355	2,2',3,5,5',6'-Hexachlorobi phenyl	52663-63-5	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-356	2,2',4,4',5,5'-Hexachlorobi phenyl	35065-27-1	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-357	2,2',4,4',6,6'-Hexachlorobi phenyl	33979-03-2	1	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-358	2,3,3',4,4',6-Hexachlorobi phenyl	74472-42-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-359	2,3',4,4',5',6-Hexachlorobi phenyl	59291-65-5	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-360	3,3',4,4',5,5'-Hexachlorobi phenyl	32774-16-6	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-361	2,2',3,3',4',5-Hexachloro-4-biphenylol	158076-62-1	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-362	2,2',3,4',5,5'-Hexachloro-4-biphenylol	145413-90-7	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-363	2',3,3',4',5,5'-Hexachloro-4-biphenylol	158076-63-2	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-497	2,2',3,3',4,4',5,5'-Octachlorobi phenyl	35694-08-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-581	2,2',3,3'-Tetrachlorobi phenyl	3844-93-8	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-582	2,2',3,4'-Tetrachlorobi phenyl	52663-59-9	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-583	2,2',3,6'-Tetrachlorobi phenyl	41464-47-5	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-585	2,2',4,5'-Tetrachlorobi phenyl	41464-40-8	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-586	2,2',4,6'-Tetrachlorobi phenyl	68194-04-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-587	2,2',5,5'-Tetrachlorobi phenyl	35693-99-3	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-588	2,2',6,6'-Tetrachlorobi phenyl	15968-05-5	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-589	2,3,3',5'-Tetrachlorobi phenyl	41464-49-7	0	1	1	0	-	-	0	0
Polychlorinated	tox-590	2,3,4,4'-	33025-41-1	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
biphenyl		Tetrachlorobiphenyl									
Polychlorinated biphenyl	tox-591	2,3',4,5'-Tetrachlorobiphenyl	73575-52-7	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-592	2,3',4',5-Tetrachlorobiphenyl	32598-11-1	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-593	2,4,4',5-Tetrachlorobiphenyl	32690-93-0	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-595	3,3',4,5-Tetrachlorobiphenyl	70362-49-1	0	1	1	0	-	-	0	0
Polychlorinated biphenyl	tox-596	2',3',5',6'-Tetrachloro-4,4'-biphenyldiol	100702-98-5	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-598	2,2',4',6'-Tetrachloro-4-biphenylol	150304-08-8	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-599	2,2',6,6'-Tetrachloro-4-biphenylol	219952-18-8	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-600	2',3',4',5'-Tetrachloro-3-biphenylol	67651-37-0	1	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-602	2',3,4',6'-Tetrachloro-4-biphenylol	189578-00-5	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-604	Tetrahydrochrysene	104460-72-2	1	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-614	2,2',5-Trichlorobiphenyl	37680-65-2	0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-615	2,4,6-Trichlorobiphenyl	35693-92-6	0	1	1	1	-	0	0	0
Polychlorinated biphenyl	tox-616	2',4',6'-Trichloro-4-biphenylol	14962-28-8	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-617	3,3',4'-Trichloro-4-biphenylol	124882-64-0	1	1	1	1	-	0	1	1
Polychlorinated biphenyl	tox-618	3,4',5-Trichloro-4-biphenylol	4400-06-0	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-240	1,6-Dimethylnaphthalene	575-43-9		1	1	0	0	0	0	
Polycyclic aromatic hydrocarbon	tox-126	Chrysene	218-01-9		1	1	0	0	0	0	
Polycyclic aromatic hydrocarbon	tox-48	Benzo[b]fluorene	243-17-4		1	1	0	0	0	0	
Polycyclic aromatic hydrocarbon	tox-238	α,α-Dimethyl-b-ethylallenolic acid	15372-37-9		1	1	1	0	0	1	
Polycyclic aromatic hydrocarbon	tox-167	Dibenz[ah]anthracene	53-70-3	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Polycyclic aromatic hydrocarbon	tox-244	5,11-trans - Dimethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-245	(5R,11R)-5,11-Dimethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-246	(5S,11S)-5,11-Dimethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-254	5,11-trans - Dipropyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-255	(5R,11R)-5,11-Dipropyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-256	(5S,11S)-5,11-Dipropyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-322	Fluoranthene	206-44-0	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-323	Fluorene	86-73-7	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-33	Anthracene	120-12-7	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-369	2-Hydroxybenzo[c]phenanthrene	22717-94-8	1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-374	2-Hydroxychrysene	65945-06-4	1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-396	2-Hydroxy-5-methylchrysene		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-397	8-Hydroxy-5-methylchrysene		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon	tox-41	Benz[a]anthracene	56-55-3	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-425	Indeno[1,2,3-cd]pyrene	193-39-5	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-43	Benzo[a]carbazole	239-01-0	0	1	1	0	-	-	0	0
Polycyclic aromatic	tox-44	Benzo[c]carbazole		0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
hydrocarbon											
Polycyclic aromatic hydrocarbon	tox-45	Benzo[b]fluoranthene	205-99-2	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-46	Benzo[k]fluoranthene	207-08-9	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-47	Benzo[a]fluorene	238-84-6	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-49	Benzo[b]naphtho[2,1-d]thiophene	239-35-0	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-50	Benzo[b]naphtho[2,3-d]thiophene	243-46-9	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-501	2,2',3,3',6-Pentachlorobiphenyl	52663-60-2	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-502	2,2',3',4',5'-Pentachloro-4-biphenylol	150304-12-4	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-503	2,2',3,4,5'-Pentachlorobiphenyl	38380-02-8	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-504	2,2',3',4',6'-Pentachloro-4-biphenylol	150304-10-2	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-505	2,2',3,4',6'-Pentachlorobiphenyl	68194-05-8	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-506	2,2',3',5',6'-Pentachloro-4-biphenylol	150304-11-3	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-507	2,2',3,5',6'-Pentachlorobiphenyl	38379-99-6	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-508	2,2',4,4',5'-Pentachlorobiphenyl	38380-01-7	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-509	2,2',4,5,5'-Pentachlorobiphenyl	37680-73-2	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-51	Benzo[ghi]perylene	191-24-2	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-510	2,2',4',6,6'-Pentachloro-4-biphenylol		1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-511	2,2',4,6,6'-Pentachlorobiphenyl	56558-16-8	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-512	2',3,3',4,4'-Pentachloro-2-biphenylol	150975-80-7	U	-	-	-	-	-	-	-
Polycyclic aromatic hydrocarbon	tox-513	2',3,3',4',5'-Pentachloro-4-biphenylol	149589-55-9	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-514	2,3,3',4',5'-Pentachloro-4-biphenylol	152969-11-4	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-515	2',3,3',4',5'-Pentachloro-4-biphenylol	192190-09-3	1	1	1	1	-	0	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Polycyclic aromatic hydrocarbon	tox-516	2',3,3',4',6'-Pentachloro-4-biphenylol	192190-10-6	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-517	2',3,3',5',6'-Pentachloro-4-biphenylol	189578-02-7	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-518	2,3,3',5,6-Pentachlorobiphenyl	74472-36-9	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-519	2,3',4,4',5-Pentachloro-3-biphenylol	170946-11-9	1	1	1	1	-	0	0	0
Polycyclic aromatic hydrocarbon	tox-52	Benzo[c]phenanthrene	195-19-7	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-520	2',3',4,4',5-Pentachloro-3-biphenylol	150975-81-8	1	1	1	1	-	0	0	0
Polycyclic aromatic hydrocarbon	tox-521	2,3,4,4',6-Pentachlorobiphenyl	74472-38-1	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-522	2',3,4',5,5'-Pentachloro-4-biphenylol	149589-56-0	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-523	3,3',4,4',5-Pentachlorobiphenyl	57465-28-8	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-524	3,3',4',5,5'-Pentachloro-4-biphenylol	130689-92-8	1	1	1	1	-	0	1	1
Polycyclic aromatic hydrocarbon	tox-528	Phenanthrene	85-01-8	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-53	Benzo[a]pyrene	50-32-8	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-54	Benzo[e]pyrene	192-97-2	1	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon	tox-563	Pyrene	129-00-0	0	1	1	0	-	-	0	0
Polycyclic aromatic hydrocarbon; Phenol	tox-205	(5R,11R)-5,11-Diethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon; Phenol	tox-206	(5S,11S)-5,11-Diethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon; Phenol	tox-216	5,11-trans-Diethyl-5,6,11,12-tetrahydrochrysene-2,8-diol		1	1	1	1	-	1	-	1
Polycyclic aromatic hydrocarbon; Phenol	tox-400	6-Hydroxytetralin	1125-78-6	0	1	1	1	-	0	1	1
Polyether	tox-74	Bisphenol A ethoxylate	68140-85-2	0	1	1	0	-	-	1	1
Pteridine	tox-325	Folic acid	59-30-3		1	1	0	0	0	0	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Purine	tox-101	Caffeine	58-08-2		1	1	0	0	0	0	
Pyrazole	tox-560	Propylpyrazol etriol		1	1	1	0	-	-	1	1
Pyrethrin; Pyrethroid	tox-261	Empenthrin	54406-48-3	0	1	0	-	1	-	0	0
Pyrethrin; Pyrethroid	tox-403	Imiprothrin	72963-72-5	0	1	0	-	1	-	0	0
Pyrethrin; Pyrethroid	tox-527	Permethrin	52645-53-1	0	1	1	0	-	-	0	0
Pyrethrin; Pyrethroid	tox-542	d -Phenothrin	26002-80-2	0	1	1	0	-	-	0	0
Pyrethrin; Pyrethroid	tox-553	Prallethrin	23031-36-9	0	1	0	-	1	-	0	0
Pyrethrin; Pyrethroid	tox-7	d-trans Allethrin	584-79-2	0	1	0	-	1	-	0	0
Resorcylic acid lactone; Phenol	tox-633	a-Zearalanol	26538-44-3		1	1	1	0	0	0	
Resorcylic acid lactone; Phenol	tox-635	Zearalanone	5975-78-0		1	1	1	0	0	1	
Resorcylic acid lactone; Phenol	tox-636	a-Zearalenol	36455-72-8		1	1	1	0	0	1	
Resorcylic acid lactone; Phenol	tox-634	b-Zearalanol	42422-68-4		1	1	1	0	0	1	
Resorcylic acid lactone; Phenol	tox-638	b-Zearalenol	71030-11-0		1	1	1	0	0	1	
Resorcylic acid lactone; Phenol	tox-637	Zearalenone	17924-92-4	1	1	1	1	-	0	1	1
Siloxane	tox-253	1,3-Diphenyltetra methylsiloxane	56-33-7		1	1	0	0	0	1	
Siloxane	tox-155	1,3-Dibenzyltetra methylsiloxane			1	1	0	0	0	0	
Steroid, nonphenolic	tox-267	Epitestosterone	481-30-1		1	0	0	1	0	1	
Steroid, nonphenolic	tox-218	5a-Dihydrotestosterone	521-18-6		1	0	0	1	0	1	
Steroid, nonphenolic	tox-554	Progesterone	57-83-0		1	0	0	1	0	1	
Steroid, nonphenolic	tox-580	Testosterone	58-22-0		1	0	0	1	0	1	
Steroid, nonphenolic	tox-24	5a-Androstane-3a,17b-diol	1852-53-5		1	0	0	1	0	1	
Steroid, nonphenolic	tox-25	5a-Androstane-3b,17b-diol	571-20-0		1	0	0	1	0	1	
Steroid, nonphenolic	tox-435	Mestranol	72-33-3		1	1	0	0	0	1	
Steroid, nonphenolic	tox-125	Cholesterol	57-88-5		1	0	0	1	0	1	
Steroid, nonphenolic	tox-5	Aldosterone	52-39-1		1	0	0	1	0	1	
Steroid, nonphenolic	tox-161	3-Deoxyestradiol	2529-64-8		1	1	0	0	0	1	
Steroid, nonphenolic	tox-166	Dexamethasone	50-02-2		1	0	0	1	0	1	
Steroid, nonphenolic	tox-134	Corticosterone	50-22-6		1	0	0	1	0	1	

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Steroid, nonphenolic	tox-573	b-Sitosterol	83-46-5		1	0	0	1	0	1	
Steroid, nonphenolic	tox-219	5b-Dihydrotestosterone	571-22-2		1	0	0	1	0	1	
Steroid, nonphenolic	tox-162	3-Deoxyestrone	53-45-2		1	1	0	0	0	1	
Steroid, nonphenolic	tox-398	16b-Hydroxy-16-methyl-17b-estradiol 3-methyl ether	3434-79-5		1	1	0	0	0	1	
Steroid, nonphenolic	tox-495	Norethynodrel	68-23-5		1	0	0	1	0	1	
Steroid, nonphenolic	tox-153	1,3-Diacetoxy-17a-ethinyl-7a-methyl-1,3,5(10)-estratrien-17b-ol		1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-154	1,3-Dibenzoyloxy-17a-ethinyl-7a-methyl-1,3,5(10)-estratrien-17b-ol		1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-157	14-Dehydroestradiol-17b 3-methyl ether	35664-58-7	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-159	14-Dehydroestrone 3-methyl ether	17550-11-7	0	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-165	Dehydroepiandrosterone	53-43-0	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-19	2-Aminoestratrien-17b-ol	17522-06-4	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-20	4-Aminoestratrien-17b-ol	17522-04-2	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-23	3b-Androstenediol	25126-76-5	1	1	0	-	1	-	1	1
Steroid, nonphenolic	tox-26	5b-Androstane-3a,17b-diol	1851-23-6	0	1	0	-	1	-	1	1
Steroid, nonphenolic	tox-27	5b-Androstenedione	5982-99-0	0	1	0	-	1	-	0	0
Steroid, nonphenolic	tox-277	17b-Estradiol 3-acetate	4245-41-4	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-278	Estradiol benzoate	50-50-0	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-279	Estradiol diacetate	3434-88-6	1	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-28	5a-Androstane-3,17-dione	846-46-8	0	1	0	-	1	-	0	0
Steroid, nonphenolic	tox-280	17b-Estradiol 3-methyl	1035-77-4	1	1	1	0	-	-	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		ether									
Steroid, nonphenolic	tox-289	Estrone 3-acetate	901-93-9	1	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-29	5a-Androstane-3a-ol-17-one	53-41-8	0	1	0	-	1	-	1	1
Steroid, nonphenolic	tox-290	Estrone 3-methyl ether	1624-62-0	0	1	0	-	1	-	0	0
Steroid, nonphenolic	tox-291	Estrone-3-sulfate	481-97-0	0	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-30	4-Androstenediol	1156-92-9	1	1	0	-	1	-	1	1
Steroid, nonphenolic	tox-31	5-Androstenediol	521-17-5	1	1	0	-	1	-	1	1
Steroid, nonphenolic	tox-317	2-Fluoroestratrie n-17b-ol	101772-22-9	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-318	4-Fluoroestratrie n-17b-ol	96607-54-4	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-32	4-Androstenedione	63-05-8	U	-	-	-	-	-	-	-
Steroid, nonphenolic	tox-420	9a-Methyl-14-dehydroestrone 3-methyl ether		0	1	1	1	-	0	1	1
Steroid, nonphenolic	tox-422	9a-Methylestradiol-17b 3-methyl ether	51242-32-1	0	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-424	9a-Methylestrone 3-methyl ether	31266-41-8	0	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-444	9a-Methyl-14-dehydroestradiol-17b 3-methyl ether	88598-64-5	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-446	7a-Methylestrone 3-methyl ether	10449-00-0	0	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-448	7a-Methylestradiol-17b 3-methyl ether	15506-01-1	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-450	7a-Methyl-14-dehydroestrone 3-methyl ether	35644-57-6	0	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-452	7a-Methyl-14-dehydroestradiol-17b 3-methyl ether	35644-59-8	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-456	3-Methoxyestriol	1474-53-9	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-461	11b-Methylestrone 3-methyl ether	13667-04-4	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Steroid, nonphenolic	tox-463	11b-Methylestradiol-17b 3-methyl ether	18046-75-8	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-465	11b-Methyl-14-dehydroestrone 3-methyl ether	88598-69-0	0	1	1	0	-	-	0	0
Steroid, nonphenolic	tox-466	11b-Methyl-14-dehydroestradiol-17b 3-methyl ether	88598-65-6	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-494	Norethindrone	68-22-4	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-496	19-Nortestosterone	434-22-0	1	1	1	0	-	-	1	1
Steroid, nonphenolic	tox-556	Promegestone	34184-77-5	1	1	0	-	1	-	0	0
Steroid, nonphenolic	tox-61	Bisdesoxyestradiol	1217-09-0	1	1	1	-	0	-	-	0
Steroid, phenolic	tox-274	17b-Estradiol	50-28-2		1	1	1	0	0	1	
Steroid, phenolic	tox-287	Estriol	50-27-1		1	1	1	0	0	1	
Steroid, phenolic	tox-288	Estrone	53-16-7		1	1	1	0	0	1	
Steroid, phenolic	tox-292	17a-Ethinyl estradiol	57-63-6		1	1	1	0	0	1	
Steroid, phenolic	tox-402	ICI 182780	129453-61-8		1	1	1	0	0	1	
Steroid, phenolic	tox-380	2-Hydroxyestradiol	362-05-0		1	1	1	0	0	1	
Steroid, phenolic	tox-381	4-Hydroxyestradiol	5976-61-4		1	1	1	0	0	1	
Steroid, phenolic	tox-476	Moxestrol	34816-55-2		1	1	1	0	0	1	
Steroid, phenolic	tox-401	ICI 164384	98007-99-9		1	1	1	0	0	1	
Steroid, phenolic	tox-273	17a-Estradiol	57-91-0		1	1	1	0	0	1	
Steroid, phenolic	tox-160	17-Desoxyestradiol	53-63-4		1	1	1	0	0	1	
Steroid, phenolic	tox-384	3-Hydroxyestra-1,3,5(10)-trien-16-one	3601-97-6		1	1	1	0	0	1	
Steroid, phenolic	tox-286	Estratriene-3,6a,17b-triol	1229-24-9		1	1	1	0	0	1	
Steroid, phenolic	tox-119	11b-Chloromethyl estradiol	71794-60-0	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-135	Cortisol	50-23-7	0	1	0	-	1	-	1	1
Steroid, phenolic	tox-139	Cycloprop[14R,15a]estra-1,3,5(10)-triene-3,17b-diol, 3',15-dihydro-	73860-54-5	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-140	Cycloprop[14S,15S]estra-1,3,5(10)-triene-3,17b-	105455-76-3	1	1	1	1	-	0	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		diol, 3',15-dihydro-									
Steroid, phenolic	tox-156	14-Dehydroestra diol-17b	58699-19-7	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-158	14-Dehydroestro ne	2119-18-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-17	2-Aminoestratri ene-3,17b-diol	107900-30-1	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-18	4-Aminoestratri ene-3,17b-diol	107900-31-2	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-233	11b-[2-(N,N - Dimethylamin o)ethoxy]estr a-1,3,5(10) triene-3,17b- diol		1	1	1	1	-	0	1	1
Steroid, phenolic	tox-237	11b-[3-(N,N - Dimethylamin o)- propoxy]estra -1,3,5 (10)- triene-3,17b- diol	130043-38-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-265	16-Epiestriol	547-81-9	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-266	17-Epiestriol	1228-72-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-268	Equilenin	517-09-9	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-269	Equilin	474-86-2	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-272	16a-Estradiol	1090-04-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-276	Estradiol 17- acetate		1	1	1	1	-	0	1	1
Steroid, phenolic	tox-281	9-dehydro- Estratetraene- 3,17b-diol	791-69-5	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-282	Estra- 1,3,5(10),6- tetraen-17- one, 3- hydroxy-		1	1	1	1	-	0	1	1
Steroid, phenolic	tox-283	Estra- 1,3,5(10)- triene-3,17b- diol, 14a,15a- epoxy-	79581-12-7	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-284	Estra- 1,3,5(10)- triene-3,17b- diol, 14b,15b- epoxy-	79645-49-1	0	1	1	1	-	0	1	1
Steroid, phenolic	tox-285	Estra- 1,3,5(10)- triene- 3,14,17b-triol	16288-09-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-293	17b-Ethynyl estradiol	4717-38-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-377	11a- Hydroxyestradi ol	1464-61-5	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-378	11b-	5444-22-4	1	1	1	1	-	0	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		Hydroxyestradiol									
Steroid, phenolic	tox-379	14b-Hydroxyestradiol	60183-66-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-382	2-Hydroxyestratrien-17b-ol	2259-89-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-383	4-Hydroxyestratrien-17b-ol	17592-89-1	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-385	2-Hydroxyestrone	362-06-1	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-413	(E)-17α-Iodovinylestradiol	82123-96-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-415	(Z)-17α-Iodovinylestradiol	177159-09-0	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-416	11-Keto-9b-estradiol		0	1	1	1	-	0	1	1
Steroid, phenolic	tox-417	16-Ketoestradiol	566-75-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-418	16α-Iodoestradiol	71765-94-1	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-419	6-Ketoestradiol	571-92-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-421	9α-Methylestradiol-17b	66463-44-3	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-423	9α-Methylestrone	71563-77-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-443	9α-Methyl-14-dehydroestronone	88598-67-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-445	9α-Methyl-14-dehydroestradiol-17b	88598-63-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-447	7α-Methylestrone	10448-96-1	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-449	7α-Methylestradiol-17b	10448-97-2	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-451	7α-Methyl-14-dehydroestronone	88958-66-7	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-453	7α-Methyl-14-dehydroestradiol-17b	88598-62-3	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-462	11b-Methylestrone	13667-06-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-464	11b-Methylestradiol-17b	23637-93-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-467	(Z)-11b-Methoxy-17α-iodovinylestradiol	177159-11-4	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-468	(E)-11b-Methoxy-17α-iodovinylestradiol	90857-55-9	1	1	1	1	-	0	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Steroid, phenolic	tox-482	2-Nitroestratrien e-3,17b-diol	6298-51-7	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-483	4-Nitroestratrien e-3,17b-diol	6936-94-3	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-484	2-Nitroestratrien -3-ol-17-one	5976-73-8	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-485	4-Nitroestratrien -3-ol-17-one	5976-74-9	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-526	Pentolame	150748-24-6	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-555	Prolame	99876-41-2	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-86	16a-Bromo-17b-estradiol	54982-79-5	1	1	1	1	-	0	1	1
Steroid, phenolic	tox-89	Butolame	150748-23-5	1	1	1	1	-	0	1	1
Stilbene	tox-221	Diethylstilbestrol	56-53-1		1	1	1	0	1	0	
Stilbene	tox-577	Tamoxifen	10540-29-1		1	1	0	0	0	1	
Stilbene	tox-163	4,4'-Diaminostilbene dihydrochloride	66635-40-3		1	1	0	0	0	0	
Stilbene	tox-242	α,α-Dimethylstilbestrol	552-80-7		1	1	1	0	1	0	
Stilbene	tox-434	Mestilbol	18839-90-2		1	1	1	0	1	0	
Stilbene	tox-222	Diethylstilbestrol dimethyl ether	130-79-0		1	1	0	0	0	1	
Stilbene	tox-130	cis - Clomiphene	15690-55-8	1	1	1	0	-	-	1	1
Stilbene	tox-131	trans - Clomiphene	911-45-5	1	1	1	0	-	-	1	1
Stilbene	tox-147	(R)-4'-Deoxyindenes trol A	138515-00-1	1	1	1	1	-	1	-	1
Stilbene	tox-148	(R)-5-Deoxyindenes trol A	138515-02-3	1	1	1	1	-	1	-	1
Stilbene	tox-151	(S)-4'-Deoxyindenes trol A	138514-99-5	1	1	1	1	-	1	-	1
Stilbene	tox-152	(S)-5-Deoxyindenes trol A	138515-01-2	1	1	1	1	-	1	-	1
Stilbene	tox-198	1,3-Diethyl-6,4'-dihydroxy-2-phenylindene		1	1	1	1	-	1	-	1
Stilbene	tox-212	3,3'-Diethylstilbestrol	5959-71-7	1	1	1	1	-	1	-	1
Stilbene	tox-223	Diethylstilbestrol epoxide	6052-82-0	1	1	1	1	-	1	-	1
Stilbene	tox-224	Diethylstilbestrol phenanthrene		1	1	1	1	-	1	1	1
Stilbene	tox-228	Dihydroxydiethyl	7507-01-9	1	1	1	1	-	1	-	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		hystilbestrol									
Stilbene	tox-405	Indanyldiethyl stilbestrol		1	1	1	1	-	1	-	1
Stilbene	tox-407	(R)-Indenestrol A	115217-03-3	1	1	1	1	-	1	-	1
Stilbene	tox-409	(S)-Indenestrol A	115217-04-4	1	1	1	1	-	1	-	1
Stilbene	tox-411	(R)-Indenestrol B		1	1	1	1	-	1	-	1
Stilbene	tox-414	(S)-Indenestrol B	115217-06-6	1	1	1	1	-	1	-	1
Stilbene	tox-570	Resveratrol	501-36-0	1	1	1	1	-	1	-	1
Stilbene	tox-59	Bis(m - acetoxy)-1,1,2-triphenylbut-1-ene	100808-56-8	1	1	1	0	-	-	1	1
Stilbene	tox-608	Tetramethylhexestrol	74385-27-6	1	1	1	1	-	1	-	1
Stilbene; Bisphenol	tox-574	4,4'-Stilbenediol	659-22-3		1	1	1	0	1	0	
Stilbene; Phenol	tox-575	4-Stilbenol	3839-46-1	0	1	1	1	-	1	-	1
Stilbene; Phenol	tox-8	p-(7-Alloxy)-11-ethylidibenzo-[b,f]thiepin-10-yl]phenol	85850-86-8	1	1	1	1	-	1	-	1
Stilbene; Piperidine; Phenol	tox-568	Raloxifene	84449-90-1	1	1	1	1	-	1	-	1
Stilbene; Triphenylethylene	tox-260	Droloxifene	82413-20-5		1	1	1	0	1	0	
Sulfoxide	tox-243	Dimethyl sulfoxide	67-68-5	0	0	-	-	-	-	-	0
Terpene	tox-481	Nerolidol	7212-44-4		0	0	0	0	0	0	
Terpene	tox-128	Cineole	470-82-6		1	0	0	1	0	0	
Tetrahydrophenanthrene	tox-258	Doisynoestrol	15372-34-6		1	1	0	0	0	1	
Thiophene	tox-368	3-Hydroxybenzo [b]naphtho[2,1-d]thiophene		1	1	1	1	-	1	-	1
Thiophene	tox-370	3-Hydroxybenzo [b]phenanthro[2,3-d]thiophene		1	1	1	1	-	1	-	1
Triazine	tox-557	Prometon	1610-18-0		1	1	0	0	0	0	
Triazine	tox-110	2-Chloro-4-amino-6-isopropylamino-1,3,5 triazine	6190-65-4	U	-	-	-	-	-	-	-
Triazine	tox-115	2-Chloro-4,6-diamino-S - triazine	3397-62-4	U	-	-	-	-	-	-	-
Triazine	tox-116	2-Chloro-4-ethylamino-6-amino-1,3,5-triazine	1007-28-9	0	1	1	0	-	-	0	0

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Triazine	tox-117	2-Chloro-4-ethylamino-6-(1-hydroxyisopropyl)amino-1,3,5-triazine	142179-80-4	0	1	1	0	-	-	0	0
Triazine	tox-118	2-Chloro-4-isopropylamino-6-(1-hydroxyisopropylamino)-1,3,5-triazine	142200-36-0	0	1	1	0	-	-	0	0
Triazine	tox-558	Propazine	139-40-2	0	1	1	0	-	-	0	0
Triazine; Aromatic amine	tox-37	Atrazine	1912-24-9		1	1	0	0	0	0	
Triphenylethylene	tox-627	Triphenylethylene	58-72-0		1	1	0	0	0	1	
Triphenylethylene	tox-399	4-Hydroxytamoxifen	68047-06-3		1	1	1	0	1	0	
Triphenylethylene	tox-132	Clomiphene citrate	50-41-9		1	1	0	0	0	1	
Triphenylethylene	tox-610	Toremifene citrate	89778-27-8		1	1	0	0	0	1	
Triphenylethylene	tox-11	3-(Alloxy)-10-ethyl-11-(4-hydroxyphenyl)dibenz[b,f]oxepin	85850-85-7	1	1	1	1	-	1	-	1
Triphenylethylene	tox-12	3-(Alloxy)-10-ethyl-11-phenyldibenz[o,b,f]thiepin	85850-82-4	1	1	1	0	-	-	1	1
Triphenylethylene	tox-124	Chlorotamoxifen	77588-46-6	0	1	1	0	-	-	1	1
Triphenylethylene	tox-13	3-(Alloxy)-11-ethyl-12-phenyl 6H-dibenzo[b,f]thiicin	85850-84-6	1	1	1	0	-	-	1	1
Triphenylethylene	tox-14	3-(Alloxy)-10-ethyl-11-phenyldibenz[b,f]oxepin	83807-07-2	1	1	1	0	-	-	1	1
Triphenylethylene	tox-15	3-(Alloxy)-11-ethyl-12-phenyl 5,6-dihydrodibenz[a,e]cyclooctene	85850-83-5	1	1	1	0	-	-	1	1
Triphenylethylene	tox-211	3-(2,3-Dihydroxypropoxy)-10-ethyl-11-phenyldibenz[b,f]oxepin	85850-89-1	1	1	1	0	-	-	1	1
Triphenylethylene	tox-217	5,6-Dihydro-8-[2-(dimethylamino)ethoxy]-12-ethyl-11-phenyldibenzo[a,e]cyclooctene,h	85850-78-8	1	1	1	0	-	-	1	1

Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		hydrate (1:4)									
Triphenylethylene	tox-234	3-[2-(Dimethylamino)ethoxy]-11-ethyl-12-phenyl]-6H-dibenzo[b,f]thioctin	85850-79-9	1	1	1	0	-	-	1	1
Triphenylethylene	tox-235	3-[2-(Dimethylamino)ethoxy]-10-ethyl-11-phenyldibenz[b,f]oxepin	85850-76-6	1	1	1	0	-	-	1	1
Triphenylethylene	tox-236	7-[2-(Dimethylamino)ethoxy]-11-ethyl-10-phenyldibenz[b,f]thiepin	85850-77-7	1	1	1	0	-	-	1	1
Triphenylethylene	tox-250	4-[1,2-(Diphenyl-1-butenyl)]phenol acetate	100808-55-7	1	1	1	0	-	-	1	1
Triphenylethylene	tox-252	4-[1-(Diphenylmethylene)propyl]phenol acetate	82333-68-4	1	1	1	0	-	-	1	1
Triphenylethylene	tox-298	3-[(10-Ethyl-11-(p-hydroxyphenyl)dibenzo[b,f]oxepin-3-yl)oxy]-1,2-propanediol, hydrate (4:1)	85850-93-7	1	1	1	1	-	1	-	1
Triphenylethylene	tox-299	3-[(10-Ethyl-11-(p-hydroxyphenyl)dibenzo[b,f]thiepin-3-yl)oxy]-1,2-propanediol	85850-94-8	1	1	1	1	-	1	-	1
Triphenylethylene	tox-300	3-[(11-Ethyl-12-(p-hydroxyphenyl)-6H-dibenzo[b,f]thiocin-3-yl)oxy]-1,2-propanediol	85864-54-6	1	1	1	1	-	1	-	1
Triphenylethylene	tox-301	3-[(6-Ethyl-5-(p-hydroxyphenyl)-11,12-dihydrodibenz[<i>a,e</i>]cycloocten-2-yl)oxy]-1,2-propanediol	85850-95-9	1	1	1	1	-	1	-	1
Triphenylethylene	tox-324	Fluorotamoxifen	73617-96-6	0	1	1	0	-	-	1	1

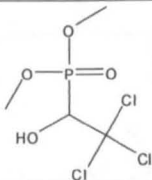
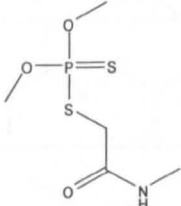
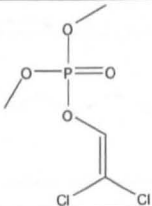
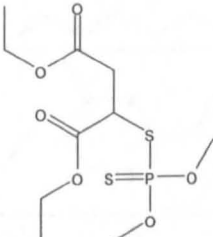
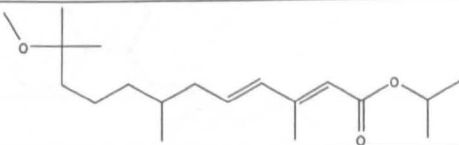
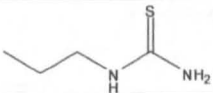
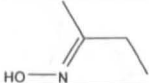
Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Triphenylethylene	tox-375	4'-Hydroxy-2,3-diphenylindenone-1		1	1	1	1	-	1	-	1
Triphenylethylene	tox-376	6'-Hydroxy-2,3-diphenylindenone-1		1	1	1	1	-	1	-	1
Triphenylethylene	tox-437	Methoxytamoxifen		1	1	1	0	-	-	1	1
Triphenylethylene	tox-440	Methyltamoxifen	73617-95-5	1	1	1	0	-	-	1	1
Triphenylethylene	tox-457	3-Methoxy-10-methyl-11-phenyldibenzo[b,f]thiepin (16b)	85807-06-1	U	-	-	-	-	-	-	-
Triphenylethylene	tox-458	2-(2-Methylphenyl)-3-phenyl-6-hydroxyindene		1	1	1	1	-	1	-	1
Triphenylethylene	tox-459	1-Methyl-6-hydroxy-2,3-diphenylindene		1	1	1	1	-	1	-	1
Triphenylethylene	tox-471	Mono-m - acetoxy-1,1,2-triphenylbut-1-ene	82333-69-5	1	1	1	0	-	-	1	1
Triphenylethylene	tox-474	Monohydroxy tamoxifen	68392-35-8	1	1	1	1	-	1	-	1
Triphenylethylene	tox-486	Nitromifene	10448-84-7	1	1	1	0	-	-	1	1
Triphenylethylene	tox-530	Phenol, 4-[3-(2dimethylamino)ethoxy]-11-ethyl-dibenzo[b,f]thioctin-12-yl]	85850-81-3	1	1	1	1	-	1	-	1
Triphenylethylene	tox-531	Phenol, 4-[7-(2-dimethylamino)ethoxy]-11-ethyl-dibenzo[b,f]thiepin-10-yl]-	85850-74-4	1	1	1	1	-	1	-	1
Triphenylethylene	tox-532	Phenol, 4-[2-(2 dimethylamino)-ethoxy]-6-ethyl-11,12-dihydro-dibenzo[a,e]cycloocten-5-yl]-	85850-75-5	1	1	1	1	-	1	-	1
Triphenylethylene	tox-533	Phenol, 3-[2-dimethylaminoethoxy]-10-ethyl- 4-hydroxyphenyl dibenzo-[b,f]oxepin	85850-80-2	1	1	1	1	-	1	-	1
Triphenylethylene	tox-534	Phenol, 4-[1-[4-[2-	96474-35-0	1	1	1	1	-	1	-	1

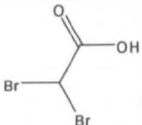
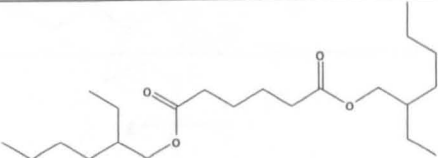
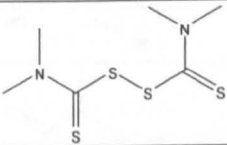
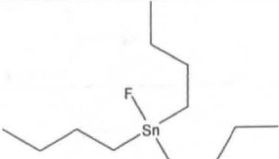
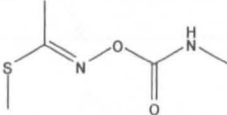
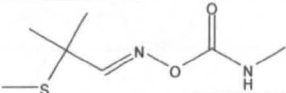
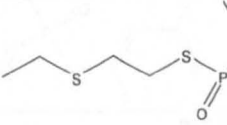
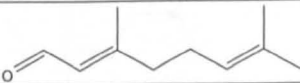
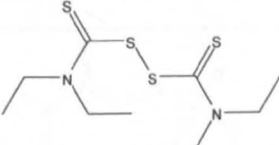
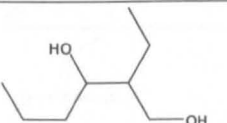
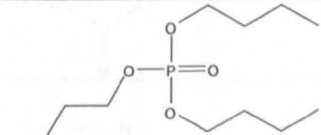
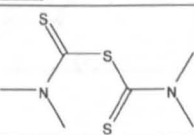
Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
		(dimethylamino)ethoxy]phenyl]-2-phenyl-1-butenyl]-3-methyl-, (E)-									
Triphenylethylene	tox-535	Phenol, 4-(1,2-diphenyl-1-butenyl)-	69967-79-9	1	1	1	1	-	1	-	1
Triphenylethylene	tox-536	Phenol, 4-(1Z)-1,2-diphenyl-1-butenyl)-	69967-80-2	1	1	1	1	-	1	-	1
Triphenylethylene	tox-537	Phenol, 4-[2-Nitro-2-phenyl-1-[4-[2-(1pyrrolidinyl)ethoxy]phenyl]ethenyl]phenyl-, (E)-	107144-84-3	1	1	1	1	-	1	-	1
Triphenylethylene	tox-538	Phenol, 4,4'-(2-phenyl-1-butenylidene)bis-	91221-46-4	1	1	1	1	-	1	-	1
Triphenylethylene	tox-564	Pyrrolidine, 1-[2-[4-[1-(4-methoxyphenyl)-2-nitro-2-phenylethenyl]phenoxy]ethyl]-, (E)	77413-87-7	1	1	1	0	-	-	1	1
Triphenylethylene	tox-60	Bis(p - acetox)-1,1,2-triphenylbut-1-ene	100808-54-6	1	1	1	0	-	-	1	1
Triphenylethylene	tox-620	Triethylamine, 2-[p-[6-methoxy-2-phenyl-3-inden-3-yl]phenoxy]hydrochloride	64-96-0	U	-	-	-	-	-	-	-
Triphenylethylene	tox-626	1,1,2-Triphenylbut-1-ene	63019-13-6	1	1	1	0	-	-	1	1
Triphenylethylene; Phenol	tox-10	p-(2-{Alloxy}-6-ethyl-11,12-dihydrodibenzo[a,e]cyclooctene-5-yl)phenol	85850-87-9	1	1	1	1	-	1	-	1
Triphenylethylene; Phenol	tox-9	p-(3-{Alloxy}-11-ethyl-6H-dibenzo[b,f]thiicin-12-yl)phenol hemihydrate	85850-88-0	1	1	1	1	-	1	-	1
Triphenylethylene; Stilbene	tox-478	Nafoxidine	1845-11-0		1	1	0	0	0	1	
Triphenylmethane	tox-539	Phenolphthal ein	77-09-8		1	1	1	0	1	0	
Triphenylmethane	tox-540	Phenolphthali n	81-90-3		1	1	1	0	1	0	

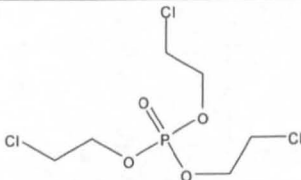
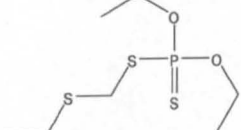

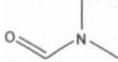

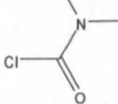
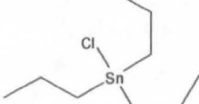
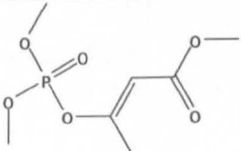
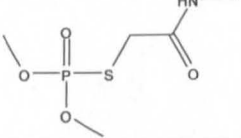
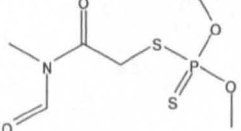
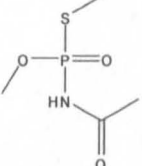
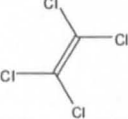
Chemical class	ID	Substance	CAS Nr.	Activity class	F1 Ring	F2 Ar	F3 Phe	F4 Het	F5 P3P	F6 Other	Predicted Class
Other	tox-297	4-Ethyl-7-hydroxy-3-(methoxyphenyl) 2H -1-benzopyran-2-one	5219-17-0		1	1	1	0	1	0	
Other	tox-311	3-[(10-Ethyl-11-phenyldibenz o-[b,f]thiepin-3-yl)oxy]-1,2-propanediol, complexed with isopropyl alcohol 2:1	85850-90-4	1	1	1	0	-	-	1	1
Other	tox-312	3-[(11-Ethyl-12-phenyl-6H -dibenzo [b,f]thiicin-3-yl)oxy]-1,2-propanediol, hydrate (4:1)	85850-92-6	1	1	1	0	-	-	1	1
Other	tox-313	3-[(6-Ethyl-5-phenyl-11,12-dihydrodibenz o [a ,e]cycloocten-2-yl)oxy]-1,2-propanediol	85850-91-5	1	1	1	0	-	-	1	1
Other	tox-566	7-Quinolinol, 1-ethyl-1,2-dihydro-3-(4-hydroxyphenyl)-4-methyl-	107144-83-2	1	1	1	1	-	1	-	1
Other	tox-567	6-Quinolinol, 1-ethyl-1,2-dihydro-3-(4-hydroxyphenyl)-4-methyl-	107144-82-1	1	1	1	1	-	1	-	1
Other	tox-613	Triaryl-pyrazole		1	1	1	0	-	-	0	0

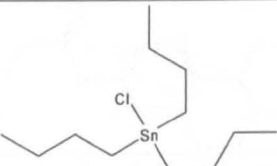
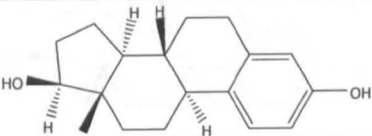
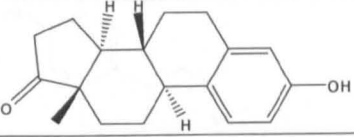
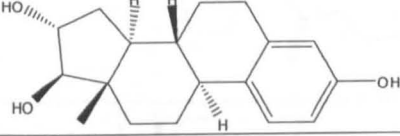
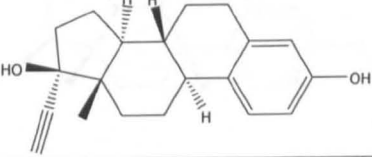
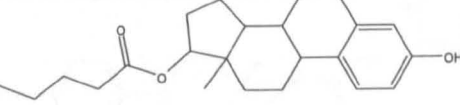
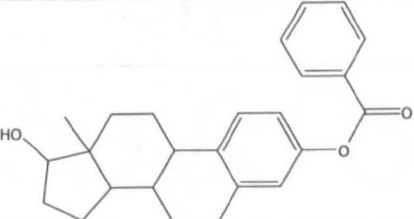
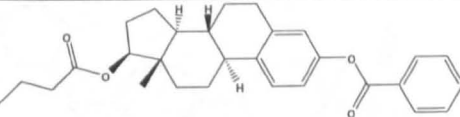
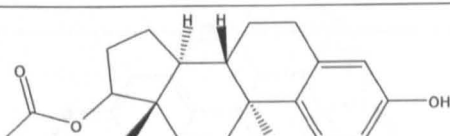
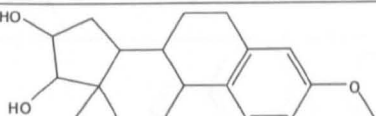
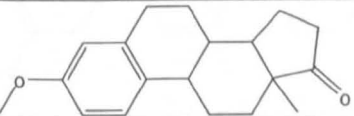
ANNEX B

Dataset extracted from the METI database [120] and used for the models developed in Chapter 5. Experimental label for both receptor binding affinity (RBA) and reporter gene assay (RA) is given (I = Inactive; A = Active) and dataset assignment is provided.

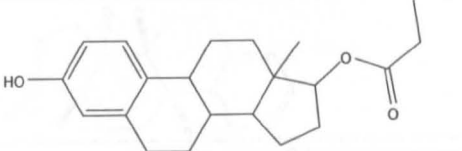
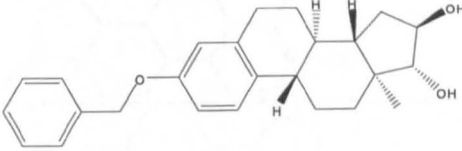
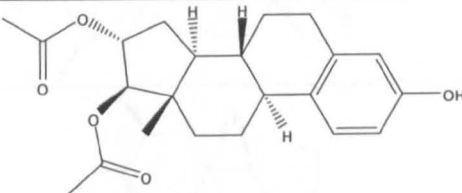
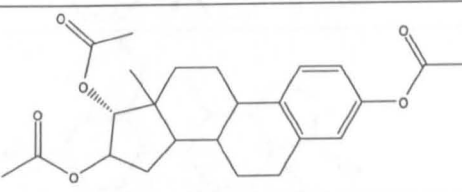
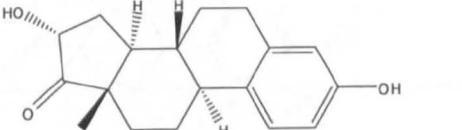
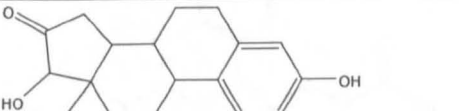
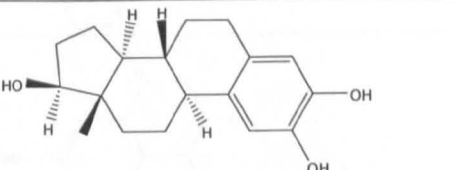
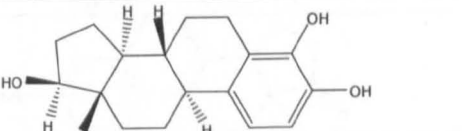
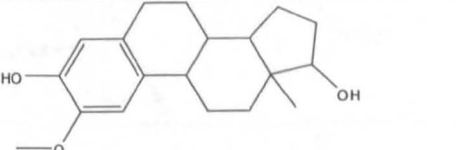

ID	Structure	CAS Nr.	Name	RBA	RA	Set
1-001		52-68-6	Trichlorfon	I	I	Test
1-002		60-51-5	Dimethoate	I	I	Train
1-003		62-73-7	Dichlorvos	I	I	Train
1-004		121-75-5	Malathion	I	I	Train
1-005		40596-69-8	Methoprene	I	I	Train
1-006		927-67-3	Propylthiourea	I	I	Train
1-007		96-29-7	2-Butanone oxime	I	I	Train

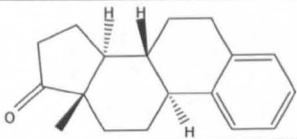

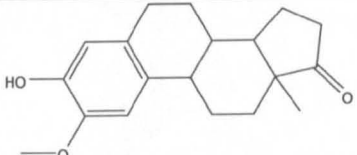
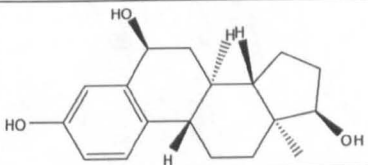
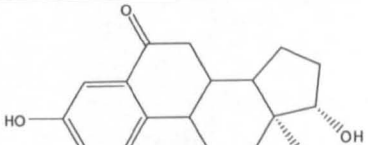
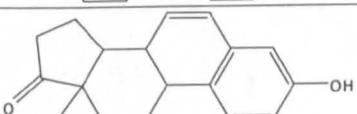
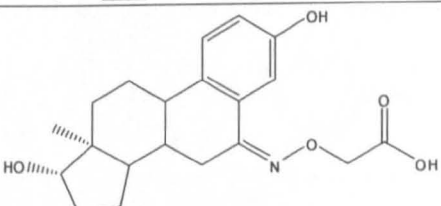
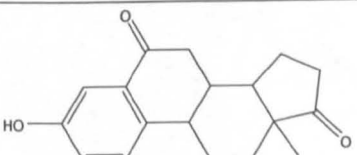
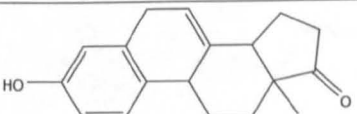
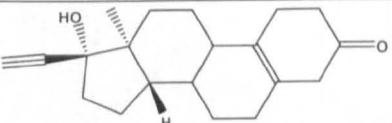

1-008		631-64-1	Dibromoacetic acid	I	I	Train
1-011		103-23-1	Bis(2-ethylhexyl)adipate	I	I	Train
1-013		137-26-8	Thiram	A	I	Val
1-015		1983-10-4	Stannane, tributylfluoro-	A	I	Train
1-016		16752-77-5	Methomyl	I	I	Val
1-017		116-06-3	Aldicarb	I	I	Test
1-018		919-86-8	Demeton-s-methyl	I	I	Test
1-019		5392-40-5	Citral	I	I	Train
1-020		97-77-8	Disulfiram	A	I	Test
1-021		94-96-2	2-Ethyl-1,3-hexanediol	I	I	Train
1-022		126-73-8	Tributyl phosphate	I	I	Train
1-023		97-74-5	Bis(dimethylthiocarbamoyl) sulfide	A	I	Train

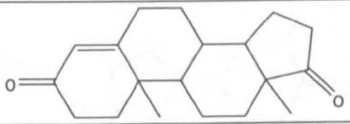

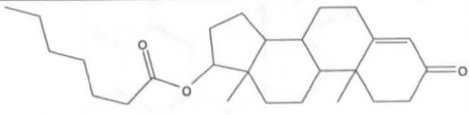
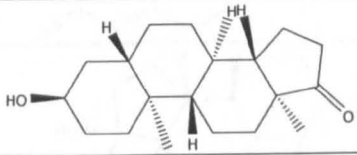
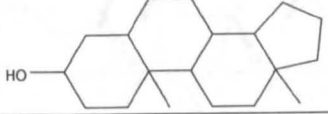
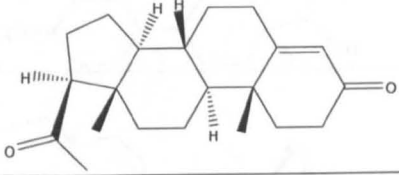
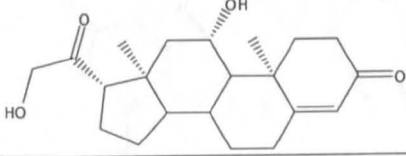
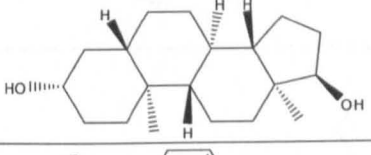

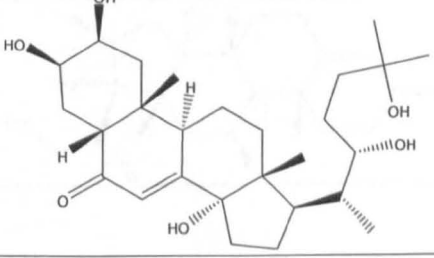
1-024		115-96-8	Tris(2-chloroethyl)phosphate	I	I	Train
1-025		298-02-2	Phorate (ISO); O,O-diethyl ethylthiomethyl phosphorodithi oate	I	I	Val
1-026		107-21-1	Ethylene glycol	I	I	Train
1-027		68-12-2	Dimethylforma mide (DMFA)	I	I	Val
1-028		96-12-8	Dibromochloro propane	I	I	Test
1-029		79-44-7	Dimethyl carbamyl chloride	I	I	Train
1-032		2279-76-7	Tri-n-propyltin (TPrT)	A	I	Train
1-033		7786-34-7	Mevinphos = Phosdrin	I	I	Val
1-036		1113-02-6	Omethoate	I	I	Val
1-037		2540-82-1	Formothion	I	I	Test
1-038		30560-19-1	Acephate	I	I	Train
1-044		127-18-4	Perchloroethyl ene	I	I	Train

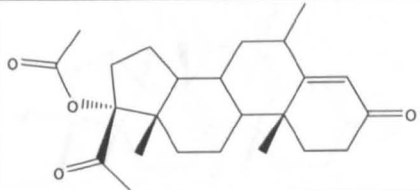
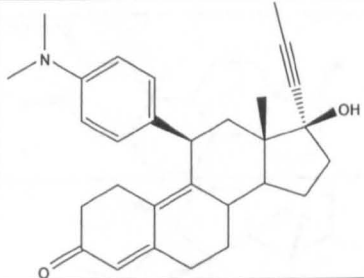
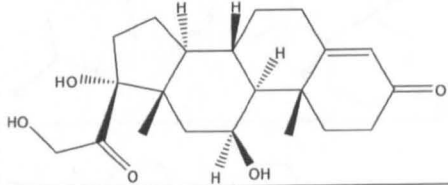

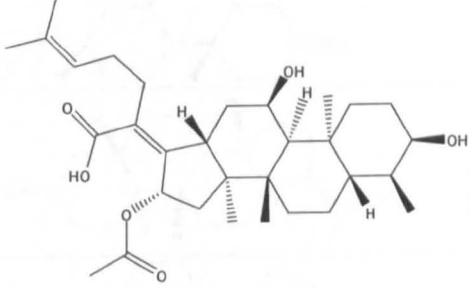
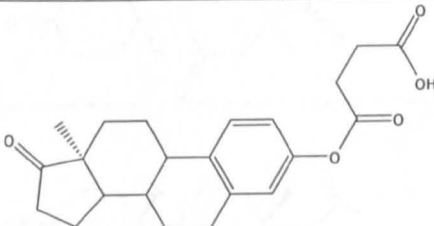
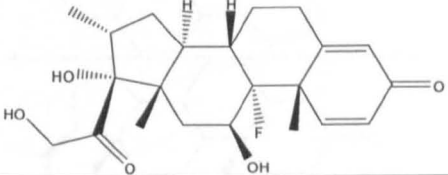
1-045		1461-22-9	Tributylchlorostannane	A	I	Val
2-001		50-28-2	beta-Estradiol	A	A	Test
2-003		53-16-7	Estrone	A	A	Train
2-004		50-27-1	Estriol	A	A	Train
2-005		57-63-6	Ethynyl estradiol	A	A	Val
2-006		979-32-8	Estradiol valerate	A	A	Train
2-007		50-50-0	beta-Estradiol-3-benzoate	A	I	Train
2-008		63042-18-2	beta-Estradiol 3-benzoate 17-n-butyrate	I	A	Train
2-009		1743-60-8	beta-Estradiol 17-acetate	A	A	Val
2-010		1474-53-9	Estriol 3-methyl ether	A	A	Train
2-011		1624-62-0	Estrone 3-methyl ether	A	A	Train

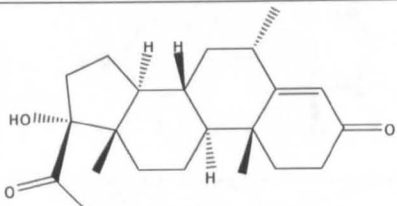
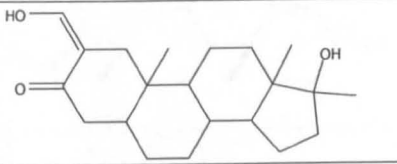
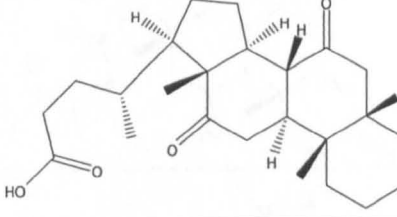
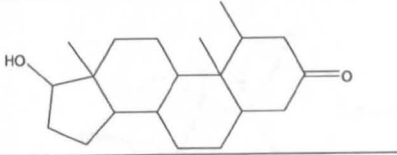
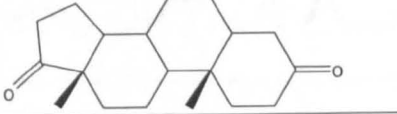
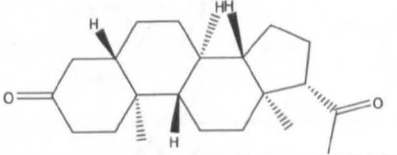
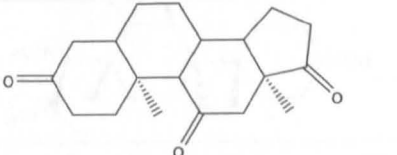
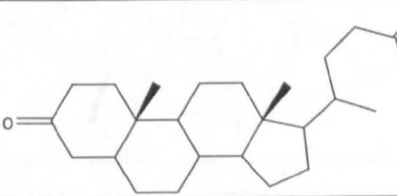

2-012		113-38-2	Estradiol dipropionate	A	A	Val
2-013		72-33-3	Mestranol	A	A	Train
2-014		313-06-4	beta-Estradiol 17-cypionate	A	A	Val
2-015		41164-36-7	beta-Estradiol 3-carboxymethyl ether	A	A	Test
2-016		7698-93-3	beta-Estradiol 17-hemisuccinate	A	A	Train
2-017		4956-37-0	beta-Estradiol 17-enanthate	A	A	Test
2-018		152-43-2	17alpha Ethynyl estradiol-3-cyclopentyl ether	A	A	Train
2-019		901-93-9	Estrone acetate	A	A	Train
2-020		69260-14-6	Estriol 3-carboxymethyl ether	A	A	Val

2-021		3758-34-7	beta-Estradiol 17-propionate	A	A	Test
2-022		18650-87-8	Estriol 3-benzyl ether	A	A	Train
2-024		805-26-5	Estriol 16,17-diacetate	A	A	Train
2-025		2284-32-4	Estriol triacetate	A	A	Train
2-026		566-76-7	16alpha-Hydroxyestrone	A	A	Val
2-027		566-75-6	16-Ketoestradiol	A	A	Train
2-028		362-05-0	2-Hydroxyestradiol	A	A	Test
2-029		5976-61-4	4-Hydroxyestradiol	A	A	Train
2-030		362-07-2	2-Methoxy-beta-estradiol	A	I	Train
2-031		1232-80-0	2-Hydroxyestriol	A	I	Train

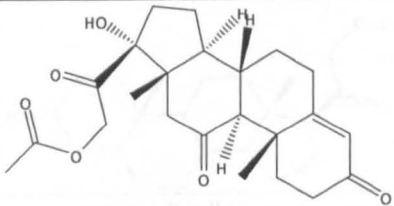
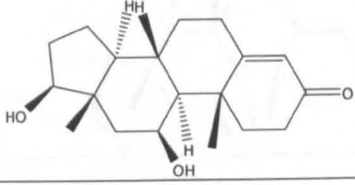
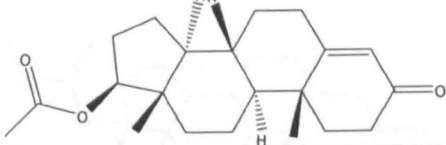
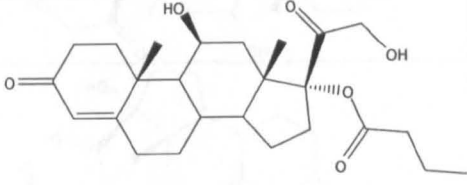
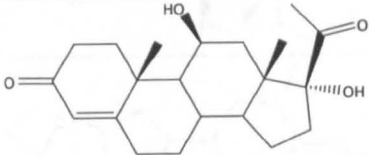
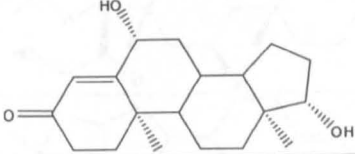
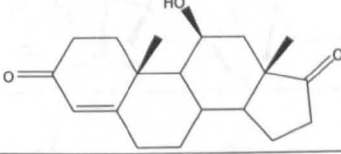
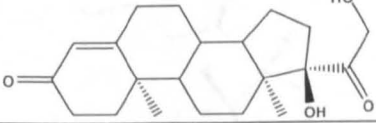
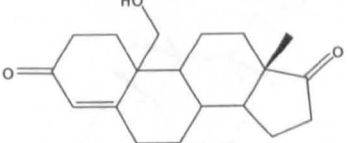
2-032		53-45-2	3-Deoxyestrone	A	A	Train
2-033		3131-23-5	4-Hydroxyestrone	A	A	Train
2-034		362-08-3	2-Methoxyestrone	A	A	Test
2-036		1229-24-9	6alpha-Hydroxyestradiol	A	A	Train
2-037		571-92-6	6-Ketoestradiol	A	A	Train
2-038		2208-12-0	6-Dehydroestrone	A	A	Test
2-039		35048-47-6	Estradiol-6-(O-carboxymethyl)oxime	A	A	Test
2-040		1476-34-2	6-Ketoestrone	A	A	Val
2-041		474-86-2	Equilin	A	A	Train
2-045		68-23-5	Norethynodrel	A	A	Train
2-046		68-22-4	Norethiridrone	A	A	Train

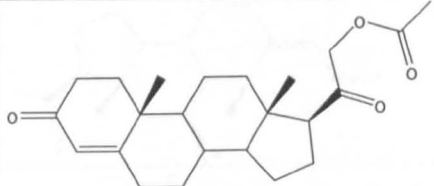
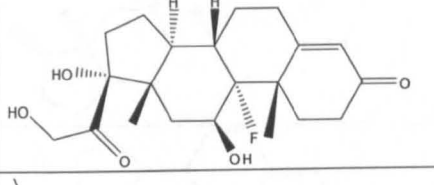
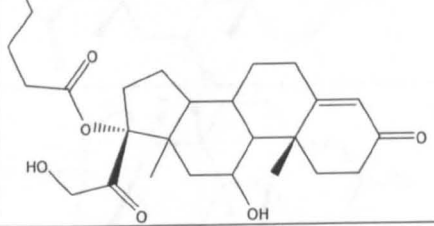
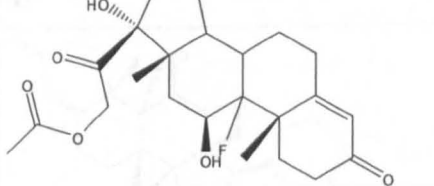
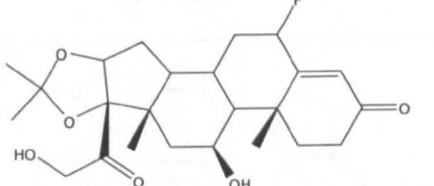
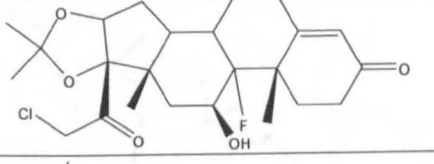
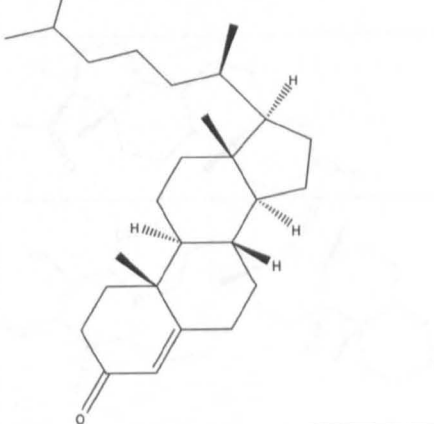
2-047		63-05-8	Androstenedione	I	I	Test
2-048		797-63-7	Levonorgestrel	A	I	Train
2-049		315-37-7	Testosterone enanthate	I	A	Train
2-050		53-41-8	Androsterone	I	I	Train
2-052		1224-92-6	5alpha-androstan-3beta-ol	A	A	Train
2-053		57-83-0	Progesterone	I	I	Train
2-054		50-22-6	(11beta)-11,21-Dihydroxypregn-4-ene-3,20-dione	I	I	Train
2-055		571-20-0	5alpha-Androstane-3beta,17beta-diol	A	A	Val
2-056		53-43-0	Dehydroepiandrosterone	A	A	Train
2-058		3604-87-3	Ecdysone	I	I	Train

2-061		71-58-9	Hydroxymethyl progesterone acetate	I	I	Test
2-062		84371-65-3	RU-486	A	I	Train
2-063		50-23-7	Cortisol	I	I	Test
2-064		434-22-0	Testosterone, 19-nor	A	A	Test
2-067		6990-06-3	Fusidic acid	I	I	Train
2-068		58534-72-8	Estrone 3-hemisuccinate	A	A	Test
2-070		50-02-2	(11beta,16alpha)-9-Fluoro-11,17,21-trihydroxy-16-methylpregna-1,4-diene-3,20-dione	I	I	Train

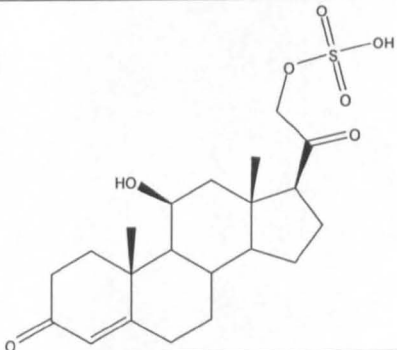
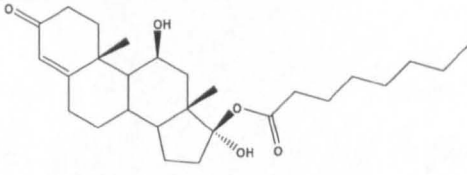
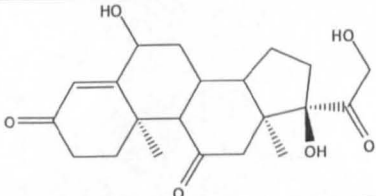
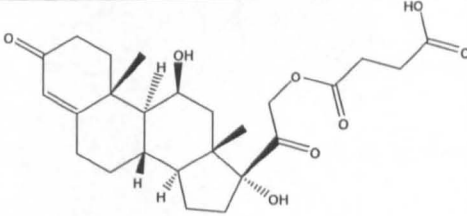
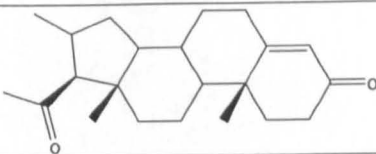
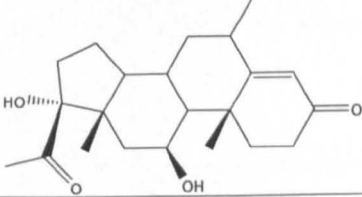
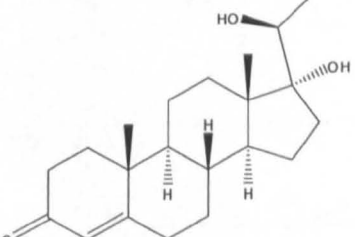
2-071		520-85-4	Progesterone, medroxy	I	I	Val
2-074		434-07-1	Oxymetholone	A	A	Train
2-075		81-23-2	Dehydrocholic acid	I	I	Train
2-076		1424-00-6	Mesterolone	A	I	Train
2-078		846-46-8	5alpha-Androstane-3,17-dione	I	I	Train
2-080		566-65-4	5alpha-Pregnane-3,20-dione	I	I	Test
2-081		1482-70-8	5alpha-Androstane-3,11,17-trione	I	I	Val
2-082		1553-56-6	5beta-Cholanic acid-3-one	I	I	Train
2-086		2089-06-7	5alpha-Pregnane-3,11,20-trione	I	I	Train

2-087		53-06-5	Cortisone	I	I	Val
2-088		64-85-7	Deoxycorticosterone	I	I	Val
2-089		50-03-3	Hydrocortisone acetate	I	I	Train
2-091		52-01-7	Spironolactone	I	I	Train
2-092		57-85-2	Testosterone propionate	I	A	Train
2-094		68-96-2	Hydroxyprogesterone	I	I	Train
2-095		302-23-8	Hydroxyprogesterone acetate	I	I	Train
2-096		630-56-8	Hydroxyprogesterone caproate	I	I	Train

2-097		50-04-4	Cortisone acetate	I	I	Test
2-098		1816-85-9	11b-Hydroxytestosterone	I	I	Val
2-099		1045-69-8	Testosterone acetate	I	I	Train
2-100		13609-67-1	Hydrocortisone-17-butyrate	I	I	Train
2-101		641-77-0	21-Deoxycortisol	I	I	Train
2-103		62-99-7	6beta-Hydroxytestosterone	I	I	Test
2-104		382-44-5	11beta-Hydroxyandrostenedione	I	I	Train
2-105		152-58-9	11-Deoxycortisol	I	I	Test
2-107		510-64-5	19-Hydroxyandrostenedione	I	I	Train

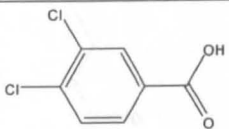
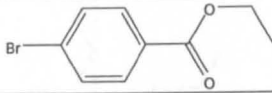
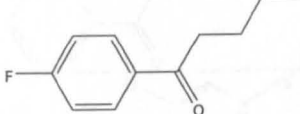
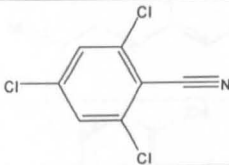
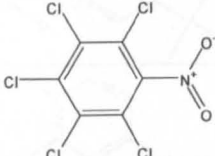
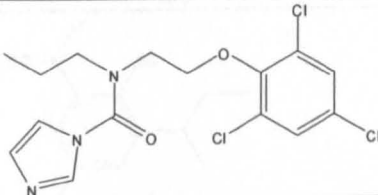
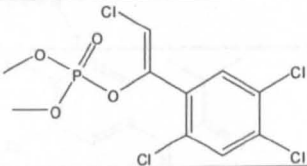
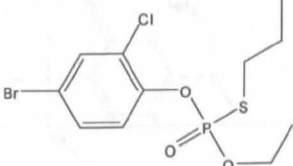
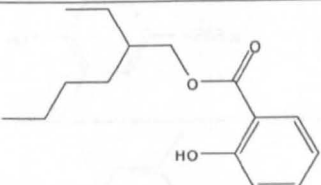
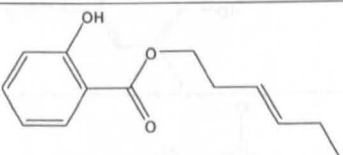
2-109		56-47-3	Deoxycorticosterone acetate	I	I	Train
2-110		127-31-1	Hydrocortisone-9a-fluoro	I	I	Train
2-111		57524-89-7	Hydrocortisone-17-valerate	I	I	Test
2-112		514-36-3	Fludrocortisone acetate	I	I	Val
2-113		1524-88-5	Fludroxycortisone	I	I	Val
2-114		3093-35-4	Halcinonide	I	I	Test
2-115		601-57-0	4-Cholesten-3-one	I	I	Train

2-116		80-75-1	11alpha-Hydroxyprogesterone	I	I	Train
2-117		382-45-6	Adrenosterone	I	I	Train
2-118		516-15-4	11-Ketoprogesterone	I	I	Test
2-120		145-14-2	20alpha-Hydroxypregn-4-en-3-one	I	I	Train
2-121		52910-82-4	d-Aldosterone 21-hemisuccinate	I	I	Val
2-123		61630-32-8	4-Androsten-4-ol-3,17-dione acetate	I	I	Val
2-125		10215-74-4	Deoxycorticosterone 21-hemisuccinate	I	I	Test
2-126		1097-51-4	16alpha,17alpha-Epoxyprogesterone	I	I	Train
2-127		4319-56-6	Deoxycorticosterone 21-glucoside	I	I	Train

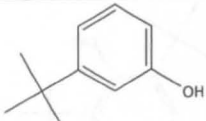
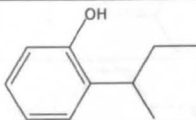
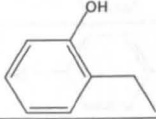
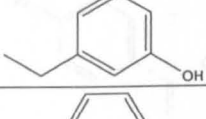
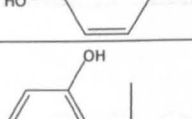
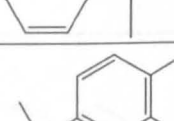
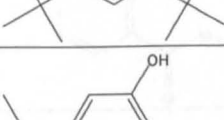
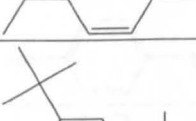
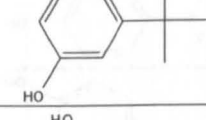
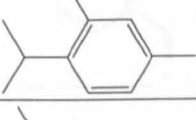
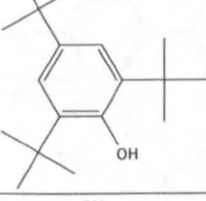
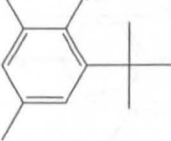
2-128		1105-02-8	Corticosterone sulfate	I	I	Train
2-129		6678-14-4	Hydrocortisone 21-caprylate	I	I	Train
2-130		16355-28-5	6beta-Hydroxycortisone	I	I	Val
2-131		2203-97-6	Hydrocortisone 21-hemisuccinate	I	I	Train
2-132		1239-79-8	16alpha-Methylprogesterone	I	I	Val
2-133		2668-66-8	6alpha-Methyl-11beta-hydroxyprogesterone	I	I	Train
2-134		1662-06-2	4-Pregnene-17alpha,20beta-diol-3-one	I	I	Test

2-135		640-87-9	Reichstein's substance 17-alpha,21-Dihydroxypregn-4-ene-3,20-dione 21-acetate	I	I	Val
2-138		15262-86-9	Testosterone isocaproate	I	I	Val
2-143		1180-25-2	Testosterone beta-D-glucuronide	I	I	Test
2-146		17230-88-5	Danazol	A	I	Val
3-001		108-95-2	Phenol	I	I	Train
3-002		123-07-9	4-Ethylphenol	I	I	Test
3-003		645-56-7	4-n-Propylphenol	I	I	Train
3-004		1638-22-8	p-Butyl phenol	A	I	Train
3-005		14938-35-3	4-n-Amylphenol	A	A	Val
3-006		2446-69-7	p-n-Hexylphenol	A	A	Train
3-007		1987-50-4	4-Heptylphenol	A	A	Test
3-009		104-43-8	4-Dodecylphenol	A	A	Val

3-010		501-24-6	3-n-Pentadecylphenol	I	I	Train
3-011		99-89-8	para-Isopropylphenol	A	I	Test
3-012		99-71-8	para-sec-Butylphenol	A	I	Train
3-013		98-54-4	p-(tert-Butyl)phenol	A	I	Val
3-014		80-46-6	p-(tert-Phenyl)phenol	A	A	Test
3-016		104-40-5	p-n-Nonylphenol	A	A	Train
3-017		140-66-9	4-tert-Octylphenol	A	A	Train
3-018		1518-83-8	4-Cyclopentylphenol	A	A	Train
3-019		1131-60-8	4-Cyclohexylphenol	A	A	Val
3-020		29799-07-3	4-(1-Adamantyl)phenol	A	A	Test
3-021		402-45-9	p-Trifluoromethylphenol	A	I	Train
3-145		41492-05-1	p-Bromobutylbenzene	I	I	Test
3-146		95-73-8	2,4-Dichlorotoluene	I	I	Train
3-147		5216-25-1	4-Chlorobenzotrichloride	I	I	Train
3-148		3972-65-4	1-Bromo-4-tert-butylbenzene	I	I	Val
3-149		50-84-0	2,4-Dichlorobenzoic Acid	I	I	Train

3-150		51-44-5	3,4-Dichlorobenzoic acid	I	I	Val
3-151		5798-75-4	Ethyl 4-bromobenzoate	I	I	Train
3-152		29114-66-7	p-Fluorovalerophenone	I	A	Train
3-153		6575-05-9	2,4,6-Trichlorobenzonitrile	I	I	Train
3-154		82-68-8	PCNB (pentachloronitrobenzene)	I	I	Val
3-155		67747-09-5	Prochloraz	I	I	Val
3-156		22248-79-9	Tetrachlorvinphos = Gardona	I	I	Train
3-157		41198-08-7	O-(4-Bromo-2-chlorophenyl)-O-ethyl-S-propyl phosphorothioate; profenofos	I	I	Train
3-249		118-60-5	2-Ethylhexyl salicylate	I	I	Train
3-251		65405-77-8	cis-3-Hexenyl salicylate	I	I	Test

3-252		56424-77-2	Ethyl Ocarboethoxy methylsalicylate	I	I	Train
3-253		93-76-5	2,4,5-T	I	I	Train
3-254		25013-16-5	Butylated hydroxyanisole	I	I	Test
3-255		95-76-1	3,4-Dichloroaniline	I	I	Test
3-256		88-85-7	Dinoseb	I	I	Train
3-257		330-54-1	Diuron	I	I	Train
3-259		21087-64-9	Metribuzin	I	I	Val
3-261		1689-84-5	Bromoxynil	I	I	Val
3-262		1689-83-4	Ioxynil	I	I	Test
3-263		94-82-6	2,4-Dichlorophenoxybutyric acid	I	I	Train

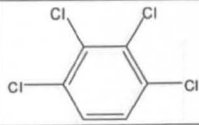
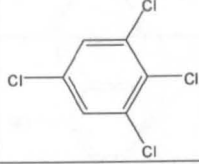
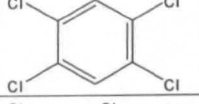
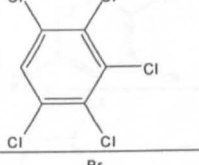
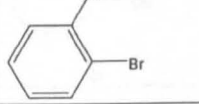
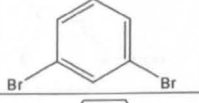
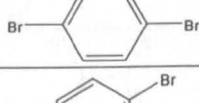
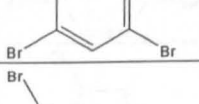
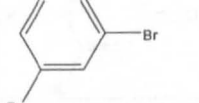
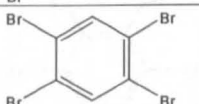
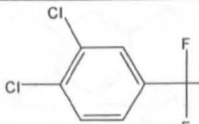
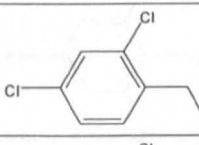

3-023		585-34-2	3-Tert-butylphenol	I	I	Val
3-024		89-72-5	ortho-sec-Butylphenol	I	I	Train
3-025		90-00-6	2-Ethylphenol	I	I	Train
3-026		620-17-7	m-Ethylphenol	I	I	Train
3-027		106-44-5	p-Cresol	I	I	Test
3-028		88-18-6	2-ter-Butylphenol	I	I	Test
3-030		96-76-4	2,4-di-tert-Butylphenol	A	I	Train
3-031		499-75-2	5-Isopropyl-2-methyl-phenol	I	I	Train
3-035		1138-52-9	Phenol, 3,5-bis(1,1-dimethylethyl)-	I	I	Test
3-036		89-83-8	Thymol	I	I	Val
3-037		732-26-3	2,4,6-Tri-t-butylphenol	I	I	Val
3-038		1879-09-0	Phenol, 2-(1,1-dimethylethyl)-4,6-dimethyl-	I	I	Train

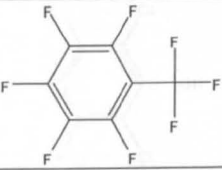
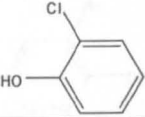
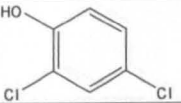
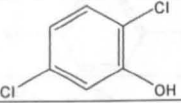
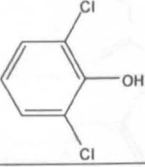
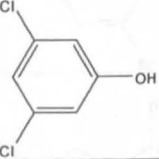
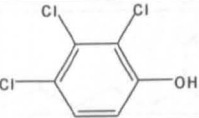
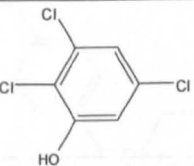
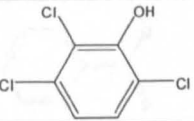
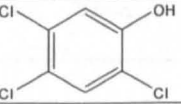
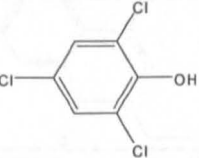
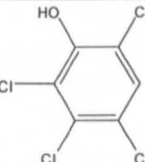
3-039		4130-42-1	Phenol, 2,6-bis(1,1-dimethylethyl)-4-ethyl-	I	I	Train
3-041		99-96-7	p-Hydroxybenzoic acid	I	I	Train
3-043		94-26-8	Butylparaben	A	A	Train
3-044		1083-27-8	Hexyl p-hydroxybenzoate	A	A	Train
3-045		2664-60-0	n-Dodecyl 4-hydroxybenzoate	A	A	Train
3-046		5153-25-3	2-Ethylhexyl 4-hydroxybenzoate	A	A	Test
3-047		94-13-3	Propylparaben	A	A	Test
3-048		6521-29-5	n-Amyl-4-hydroxybenzoate	A	A	Train
3-049		94-18-8	Benzyl-4-hydroxybenzoate	A	A	Train
3-050		4191-73-5	Isopropyl-4-hydroxybenzoate	A	A	Train
3-051		6521-30-8	4-Hydroxybenzoic acid isoamyl ester	A	A	Val
3-052		123-31-9	Hydroquinone	I	I	Test
3-056		622-62-8	4-Ethoxyphenol	I	I	Train

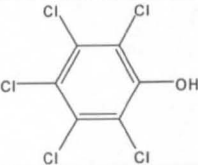
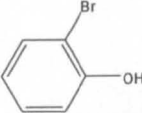
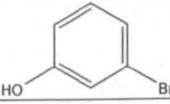

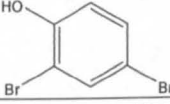
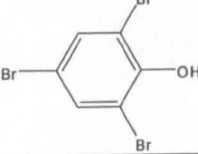
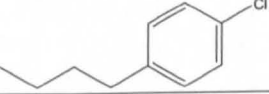
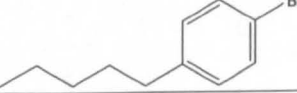
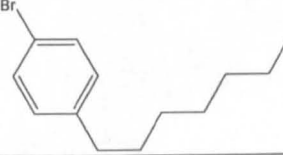
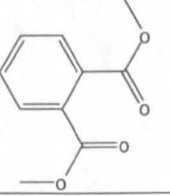
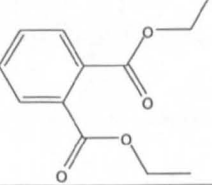
3-057		18979-50-5	4-Propoxyphenol	I	I	Train
3-058		18979-53-8	p-(Pentyloxy)phenol	A	A	Train
3-061		24305-56-4	4-n-Dodecylresorcinol	A	I	Train
3-062		98-29-3	4-tert-Butylcatechol	A	I	Train
3-063		500-66-3	Olivetol	I	I	Train
3-064		2138-20-7	4-Cyclohexylresorcinol	A	A	Train
3-065		63177-57-1	Dihydroxycinnamic acid methyl ester	A	I	Train
3-067		3943-89-3	Ethyl 3,4-dihydroxybenzoate	I	I	Val
3-070		1034-01-1	Octyl gallate	A	I	Train
3-071		1138-60-9	Isopropyl gallate	A	I	Train
3-072		84375-71-3	Hexyl vanillate	I	A	Train
3-073		2524-37-0	Ethyl 2,4-dihydroxy-6-methylbenzoate	A	I	Val

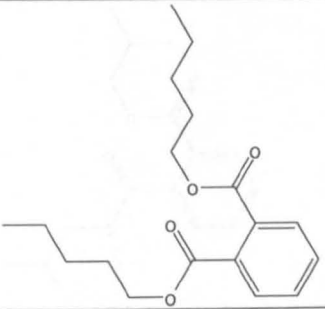
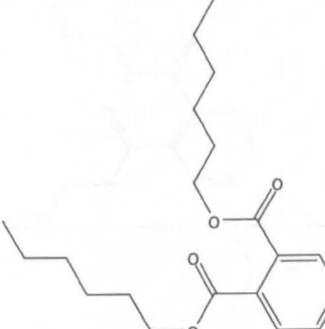

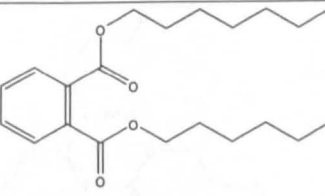
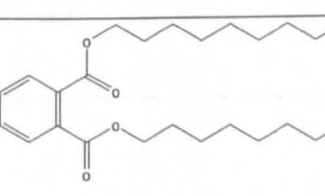
3-074		6259-76-3	Hexyl salicylate	I	I	Val
3-076		3337-59-5	Methyl 3,5-dichloro-4-hydroxybenzoate hemihydrate	I	I	Train
3-077		10210-17-0	3-(4-Hydroxyphenyl)-1-propanol	A	I	Train
3-079		3144-54-5	4-Hexanoylresorcinol	A	A	Train
3-080		70-70-2	4-Hydroxypropio phenone	I	I	Train
3-081		7400-08-0	p-Coumaric acid (PCA)	I	I	Train
3-083		14392-69-9	4'-Hydroxynonanophenone	A	A	Test
3-084		5597-50-2	Methyl 3-(4-hydroxyphenyl)propionate	I	I	Train
3-085		17362-17-3	3-(4-Hydroxyphenyl)propionitrile	I	I	Val
3-086		59-50-7	4-Chloro-3-methylphenol	I	I	Train
3-087		1570-64-5	4-Chloro-o-cresol	I	I	Train
3-088		6640-27-3	2-Chloro-4-methylphenol	I	I	Test
3-089		88-04-0	4-Chloro-3,5-xenol	I	I	Train

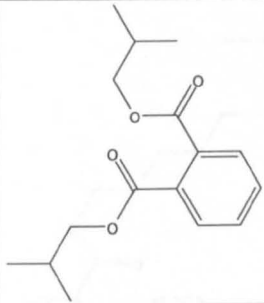
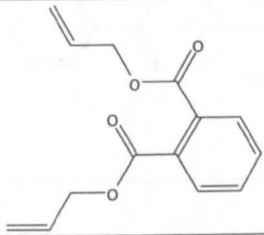
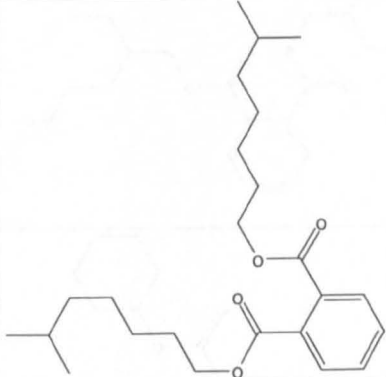
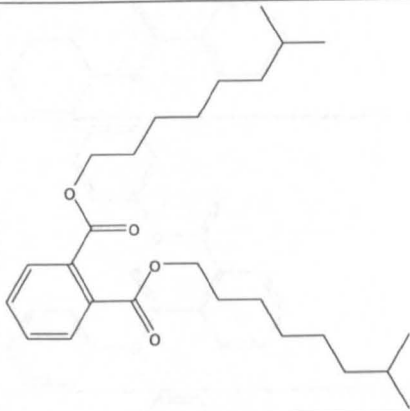
3-090		1124-04-5	2-Chloro-4,5-dimethylphenol	I	I	Train
3-092		576-55-6	3,4,5,6-Tetrabromocresol	I	I	Train
3-094		80866-89-3	5-Ethyl-5-(4-hydroxyphenyl)barbituric acid	I	I	Train
3-095		615-74-7	6-Chloro-m-cresol	I	I	Val
3-096		14199-15-6	Methyl 4-hydroxyphenyl acetate	A	I	Test
3-097		501-97-3	Phloretic acid	I	I	Train
3-098		5471-51-2	4-(4-Hydroxyphenyl)-2-butanone	I	I	Test
3-100		56718-71-9	4-(2-Methoxyethyl)phenol	I	I	Test
3-102		108-90-7	Chlorobenzene	I	I	Train
3-103		106-46-7	para-Dichlorobenzene	I	A	Val
3-104		541-73-1	m-Dichlorobenzene	I	I	Train
3-105		95-50-1	o-Dichlorobenzene	I	I	Val
3-106		120-82-1	1,2,4-Trichlorobenzene	I	I	Train
3-107		108-70-3	1,3,5-Trichlorobenzene	I	I	Train

3-108		634-66-2	1,2,3,4-Tetrachlorobenzene	A	I	Train
3-109		634-90-2	1,2,3,5-Tetrachlorobenzene	I	I	Train
3-110		95-94-3	1,2,4,5-Tetrachlorobenzene	I	I	Val
3-111		608-93-5	Pentachlorobenzene	I	I	Test
3-113		583-53-9	1,2-Dibromobenzene	I	I	Train
3-114		108-36-1	1,3-Dibromobenzene	I	I	Train
3-115		106-37-6	1,4-Dibromobenzene	I	I	Test
3-116		615-54-3	1,2,4-Tribromobenzene	I	I	Train
3-117		626-39-1	1,3,5-Tribromobenzene	I	A	Test
3-118		636-28-2	1,2,4,5-Tetrabromobenzene	I	I	Val
3-119		328-84-7	1,2-Dichloro-4-(trifluoromethyl)benzene	I	I	Val
3-120		94-99-5	2,4-Dichlorobenzyl chloride	I	I	Train
3-121		320-60-5	2,4-Dichloro-1-(trifluoromethyl)benzene	I	I	Test


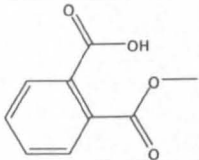
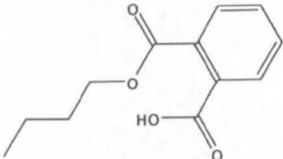
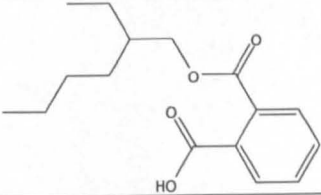
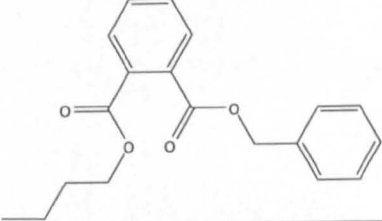
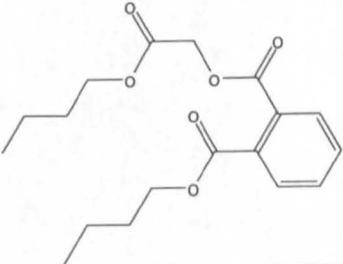
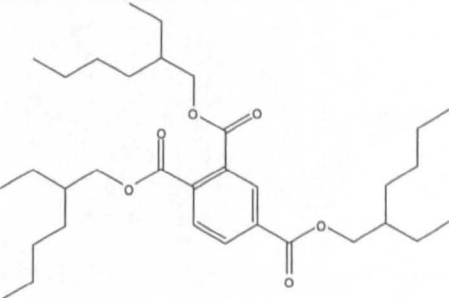
3-122		434-64-0	Perfluorotoluene	I	I	Train
3-123		95-57-8	o-Chlorophenol	I	I	Val
3-125		120-83-2	2,4-Dichlorophenol	I	I	Train
3-126		583-78-8	2,5-Dichlorophenol	I	A	Train
3-127		87-65-0	2,6-Dichlorophenol	I	I	Test
3-128		591-35-5	3,5-Dichlorophenol	I	I	Train
3-130		15950-66-0	2,3,4-Trichlorophenol	I	I	Val
3-131		933-78-8	2,3,5-Trichlorophenol	I	I	Train
3-132		933-75-5	2,3,6-Trichlorophenol	I	I	Val
3-133		95-95-4	2,4,5-Trichlorophenol	I	I	Train
3-134		88-06-2	2,4,6-Trichlorophenol	I	I	Test
3-135		58-90-2	2,3,4,6-Tetrachlorophenol	I	I	Train

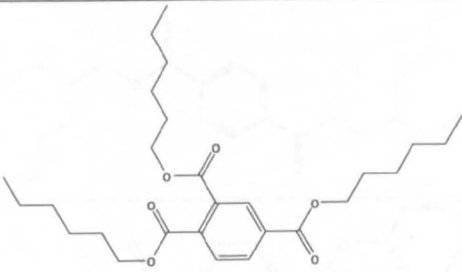
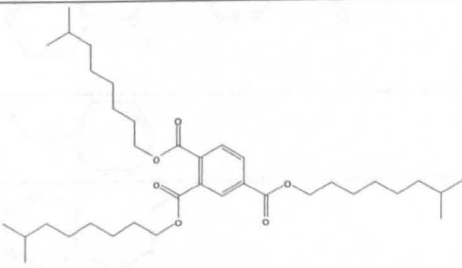
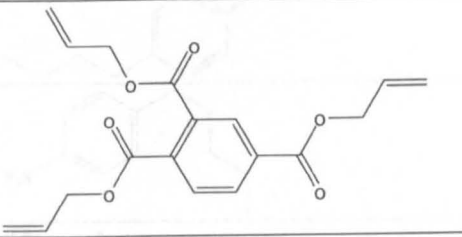
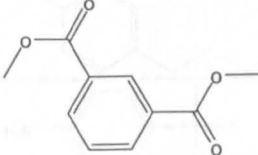
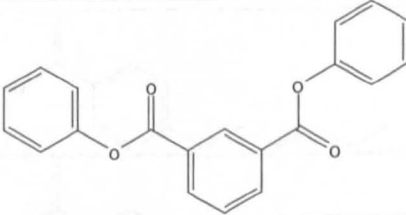
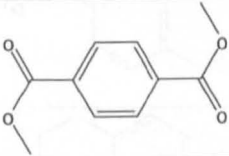
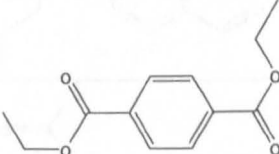
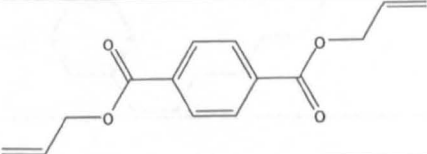
3-136		87-86-5	Pentachlorophenol	I	I	Test
3-137		95-56-7	2-Bromophenol	I	I	Train
3-138		591-20-8	3-Bromophenol	I	I	Train
3-139		106-41-2	4-Bromophenol	I	I	Val
3-140		615-58-7	2,4-Dibromophenol	I	I	Train
3-141		118-79-6	2,4,6-Tribromophenol	I	I	Train
3-142		15499-27-1	4-n-Butylchlorobenzene	I	I	Train
3-143		51554-95-1	4-n-Amylbromobenzene	I	I	Val
3-144		76287-49-5	1-Bromo-4-n-heptylbenzene	I	I	Train
3-158		131-11-3	Dimethylphthalate	I	I	Val
3-159		84-66-2	Diethylphthalate	I	I	Test

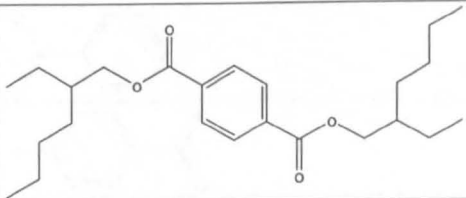
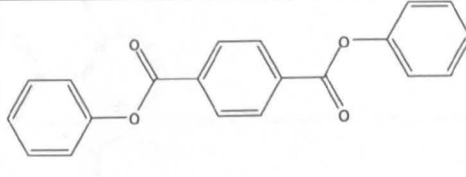
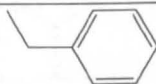
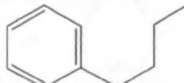
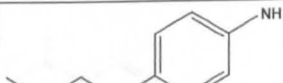
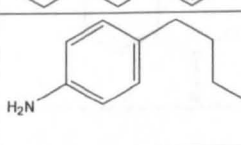
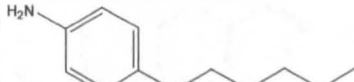
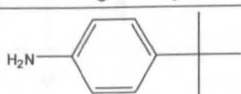
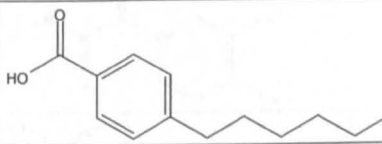
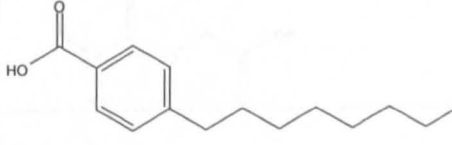
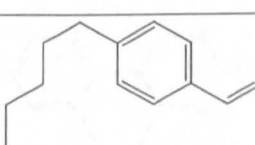
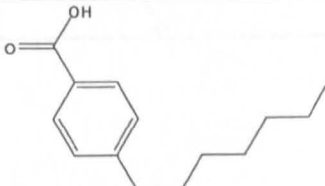
3-162		131-18-0	di-N-Pentylphthalate	A	A	Train
3-163		84-75-3	di(N-hexyl)Phthalate	A	I	Test
3-164		3648-21-3	Diheptyl phthalate	A	I	Train
3-167		84-77-5	Didecyl phthalate	I	I	Val
3-168		2432-90-8	1,2-Benzenedicarboxylic acid, didodecyl ester	I	I	Train

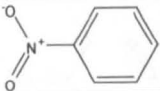
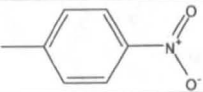
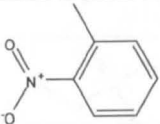
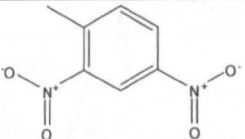
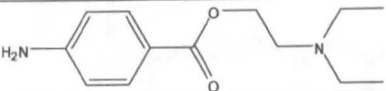
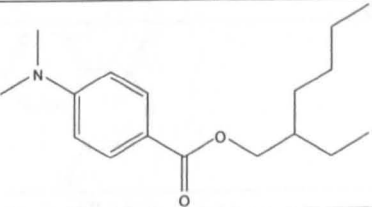
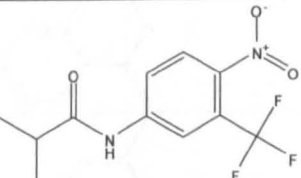
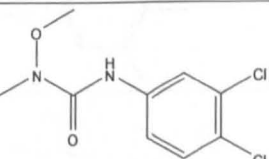
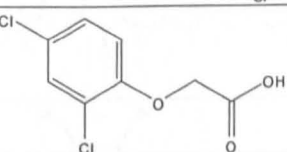
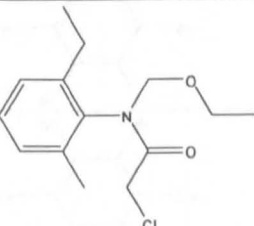
3-170		84-69-5	Diisobutyl phthalate	A	I	Test
3-171		131-17-9	Diallyl phthalate	I	I	Train
3-172		27554-26-3	1,2-Benzenedicarboxylic acid diisooctyl ester	A	I	Val
3-173		28553-12-0	Diisononyl phthalate	A	I	Test

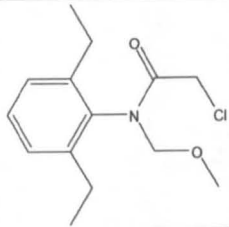
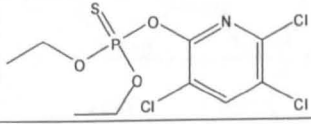
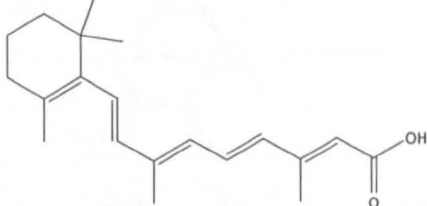
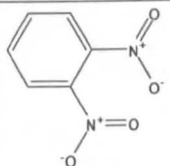
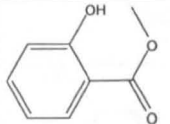
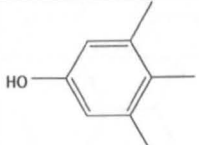
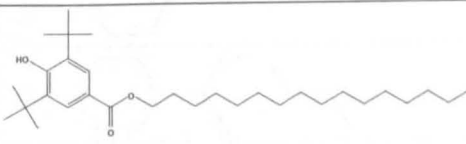
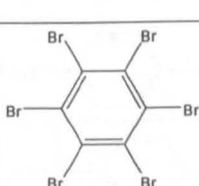
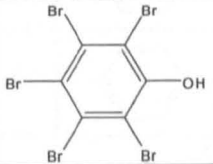
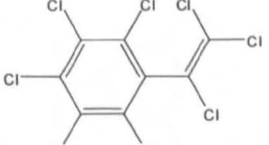
3-174		26761-40-0	1,2-Benzenedicarboxylic acid diisodecyl ester	A	I	Val
3-175		117-81-7	di-sec-Octyl phthalate	A	I	Train
3-176		84-61-7	Dicyclohexyl phthalate	A	I	Train
3-177		84-62-8	Diphenyl phthalate	I	I	Train
3-178		523-31-9	Dibenzyl phthalate	A	A	Train

3-179		117-83-9	bis(2-N-Butoxyethyl)-phthalate	I	A	Train
3-180		4376-18-5	mono-Methyl phthalate	I	I	Train
3-181		131-70-4	mono-n-Butylphthalate	I	I	Val
3-182		4376-20-9	Phthalate, monoethylhexyl	A	I	Train
3-183		85-68-7	Butylbenzyl phthalate	A	A	Train
3-184		85-70-1	Butylphthalyl butylglycolate	I	I	Test
3-185		3319-31-1	1,2,4-Benzenetricarboxylic acid tris(2-ethylhexyl) ester	I	I	Test

3-186		1528-49-0	Trihexyl trimellitate	I	I	Train
3-187		53894-23-8	Triisononyl trimellitate	I	I	Train
3-188		2694-54-4	Triallyl trimellitate	I	I	Val
3-189		1459-93-4	1,3-Benzenedicarboxylic acid, dimethyl ester	I	I	Train
3-190		744-45-6	Diphenyl isophthalate	I	A	Test
3-191		120-61-6	Dimethyl terephthalate	I	I	Val
3-192		636-09-9	Diethyl terephthalate	I	I	Train
3-193		1026-92-2	Diallyl terephthalate	I	A	Train

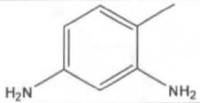
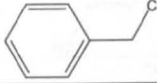
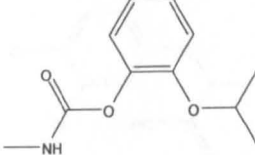
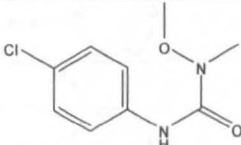
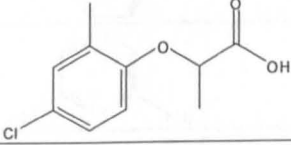
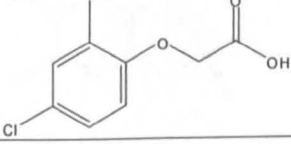
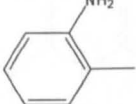
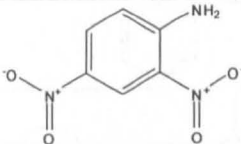
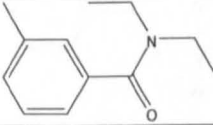
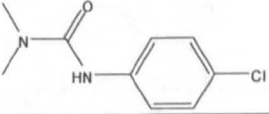
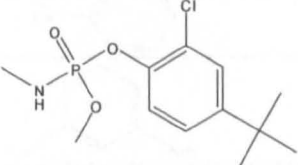
3-194		6422-86-2	Terephthalic acid bis(2-ethylhexyl)ester	I	I	Train
3-195		1539-04-4	Diphenyl terephthalate	I	A	Train
3-196		100-41-4	Ethyl benzene	I	I	Val
3-197		104-51-8	Butyl benzene	I	I	Train
3-198		104-13-2	4-Butylaniline	I	I	Train
3-199		33228-44-3	4-Pentylbenzenamine	I	I	Train
3-200		33228-45-4	4-N-Hexylaniline	I	I	Test
3-201		769-92-6	4-Tert-butylaniline	I	I	Train
3-202		21643-38-9	4-Hexylbenzoic acid	I	I	Test
3-203		3575-31-3	4-Octylbenzoic acid	I	I	Val
3-204		6853-57-2	Benzaldehyde, 4-pentyl-	I	I	Val
3-205		38350-87-7	4-Heptylbenzoic acid	I	I	Train

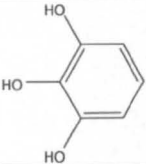
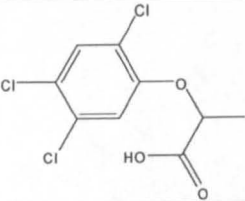
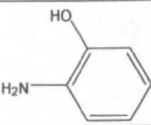
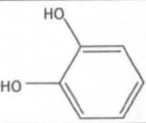
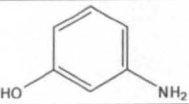
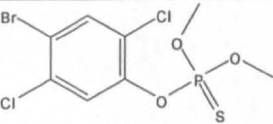
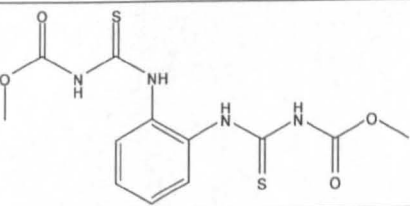
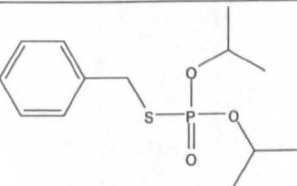
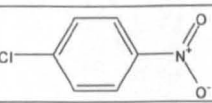
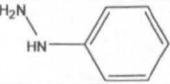
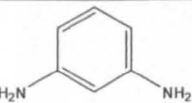
3-206		98-95-3	Nitrobenzene	I	I	Train
3-207		99-99-0	para-Nitrotoluene	I	I	Train
3-208		88-72-2	2-Nitrotoluene	I	I	Val
3-209		121-14-2	2,4-Dinitrotoluene	I	I	Train
3-210		51-05-8	Procaine hydrochloride	I	I	Train
3-211		21245-02-3	2-Ethylhexyl-4-dimethylamino benzoate	I	I	Train
3-212		13311-84-7	Flutamide	I	I	Train
3-213		330-55-2	Linuron = Lorox	I	I	Train
3-215		94-75-7	2,4-Dichlorophenoxyacetic acid	I	I	Train
3-216		34256-82-1	Acetochlor	I	I	Val

3-217		15972-60-8	Alachlor	I	I	Train
3-218		2921-88-2	Chlorpyrifos	A	I	Val
3-221		5300-03-8	Retinoic acid	I	I	Train
3-223		528-29-0	1,2-Dinitrobenzene	I	I	Test
3-224		119-36-8	Methyl salicylate	I	I	Train
3-225		527-54-8	3,4,5-Trimethylphenol	I	I	Train
3-226		67845-93-6	3,5-bis[1,1-Dimethylethyl]-4-hydroxybenzoic acid hexadecyl ester	I	I	Train
3-228		87-82-1	Hexabromobenzene	I	I	Test
3-229		608-71-9	Pentabromophenol	I	I	Train
3-232		29082-74-4	Octachlorostyrene	I	I	Train

3-233		117-82-8	1,2-Benzenedicarboxylic acid, bis(2-methoxyethyl) ester	I	I	Train
3-234		18144-43-9	4-Aminobenzoic acid 1-methylethyl ester	I	I	Test
3-235		451-46-7	4-Fluorobenzoic acid ethyl ester	I	I	Train
3-236		99-77-4	4-Nitrobenzoic acid ethyl ester	I	I	Train
3-237		16245-79-7	4-Octylbenzamine	I	I	Train
3-238		2033-89-8	3,4-Dimethoxyphenol	I	I	Train
3-239		13608-87-2	2',3',4'-Trichloroacetophenone	I	I	Test
3-240		94-25-7	n-Butyl-p-aminobenzoate	A	I	Train
3-241		94-12-2	p-Aminobenzoic acid, propyl ester	I	I	Train
3-242		455-24-3	p-Trifluoromethylbenzoic acid	I	I	Train
3-243		455-19-6	alpha,alpha,alpha-Trifluoro-p-tolualdehyde	I	I	Train
3-244		349-58-6	3,5-bis(Trifluoromethyl)phenol	I	I	Train

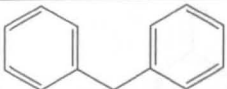
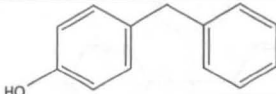
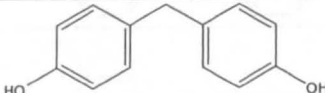
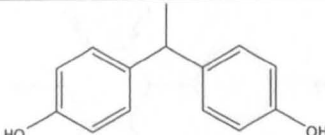
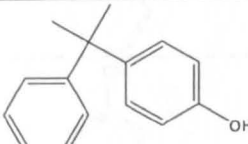
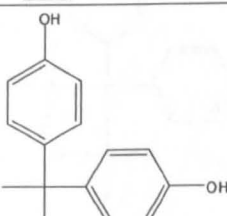
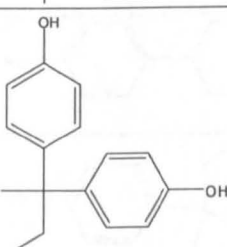
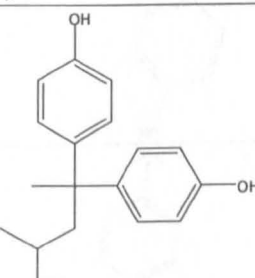
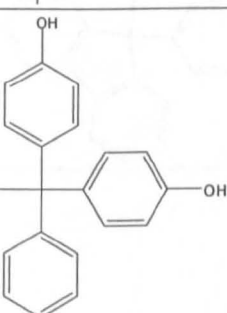
3-245		120-48-9	Butyl 4-nitrobenzoate	I	I	Train
3-246		828-27-3	4-(Trifluoromethoxy)phenol	A	I	Train
3-247		13036-02-7	Dimethyl 5-hydroxyisophthalate	I	I	Train
3-248		N.A.	Isoamyl 4-(dimethylamino)benzoate	I	I	Train
3-264		4342-36-3	Stannane, (benzoyloxy)tri n-butyl-	A	I	Test
3-265		537-98-4	Ferulic acid (FA)	I	I	Train
3-268		23950-58-5	Pronamide	I	I	Train
3-270		55-38-9	Fenthion	I	I	Val
3-272		123-30-8	4-Aminophenol	A	I	Train
3-274		299-84-3	Ronnel	A	I	Train
3-275		120-36-5	Dichloroprop	I	I	Train

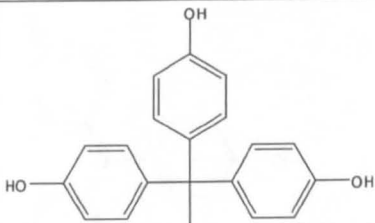
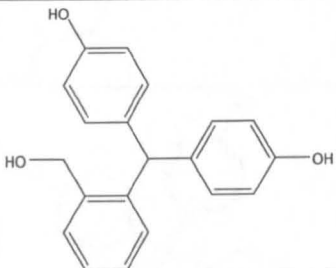
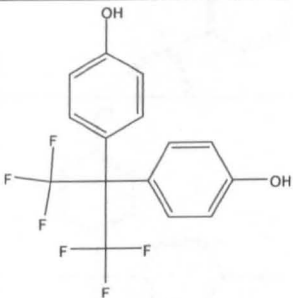
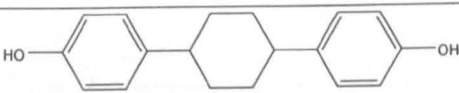
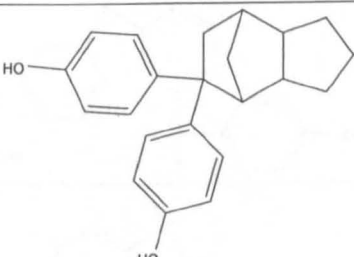
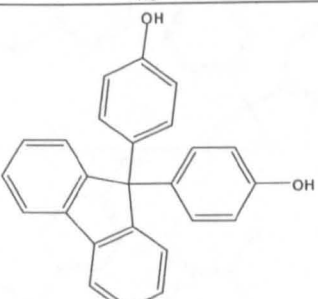
3-276		95-80-7	Toluene-2,4-diamine	I	I	Train
3-277		100-44-7	Benzyl chloride	I	I	Train
3-278		114-26-1	Propoxur(ISO) ; 2-isopropoxyphenyl N methylcarbamate; 2-isopropoxyphenyl methylcarbamate	I	I	Test
3-279		1746-81-2	Monolinuron (ISO) ; 3-(4-chlorophenyl)-1-methoxy-1-methylurea	I	I	Train
3-281		93-65-2	MCPP	I	I	Train
3-282		94-74-6	Mcpa	I	I	Test
3-283		95-53-4	ortho-Toluidine	I	I	Test
3-284		97-02-9	2,4-Dinitroaniline	I	I	Val
3-285		134-62-3	N,N-Diethyl-3-methylbenzamide	I	I	Train
3-286		150-68-5	Monuron	I	I	Test
3-287		299-86-5	Crufomat (ISO) ; 4-tert-butyl-2-chlorophenyl methyl methylphosphoramidate	I	I	Train

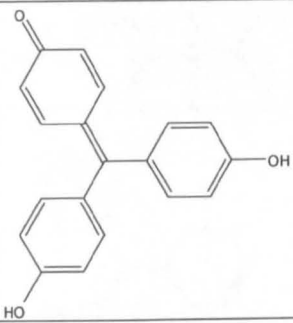
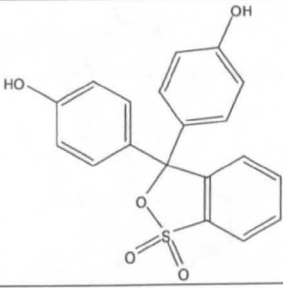
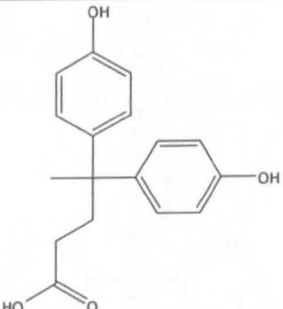
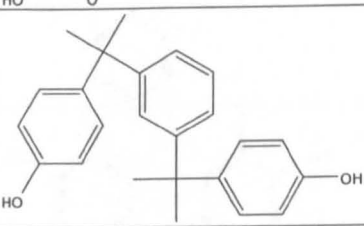
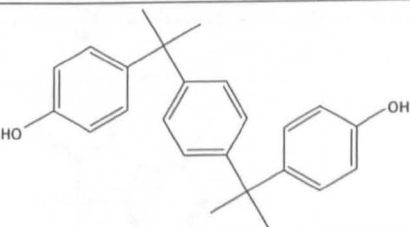
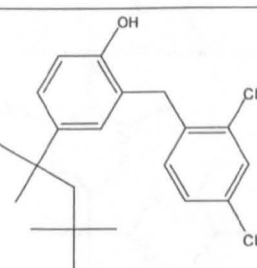
3-288		87-66-1	Pyrogallol	A	I	Train
3-289		93-72-1	Silvex	I	I	Val
3-290		95-55-6	2-Aminophenol	I	I	Train
3-291		120-80-9	Pyrocatechol	I	I	Test
3-292		591-27-5	3-Aminophenol	I	I	Train
3-293		2104-96-3	Bromophos (ISO) ; O-4-bromo-2,5-dichlorophenyl O,O-dimethyl phosphorothioate	A	I	Test
3-294		23564-05-8	Thiophanate-methyl; 1,2-di-(3-methoxycarbonyl-2-thioureido)benzene	I	I	Val
3-295		26087-47-8	S-Benzyl diisopropyl phosphorothioate; iprobenfos	I	I	Train
3-297		100-00-5	p-Nitrochlorobenzene	I	I	Train
3-298		100-63-0	Phenylhydrazine	I	I	Test
3-300		108-45-2	meta-Phenylenediamine	I	I	Val

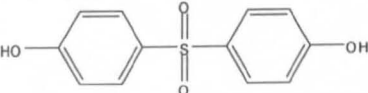
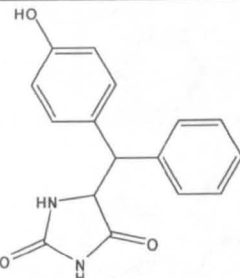
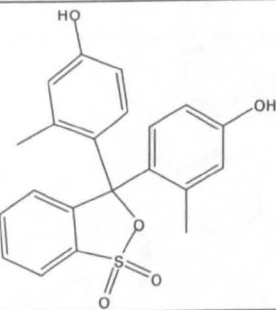
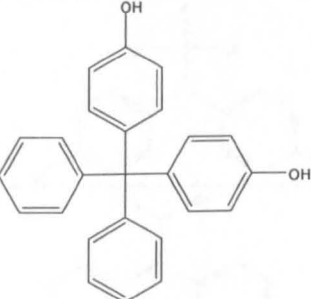
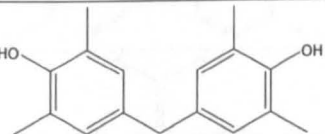
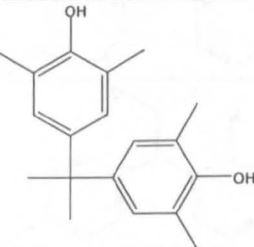
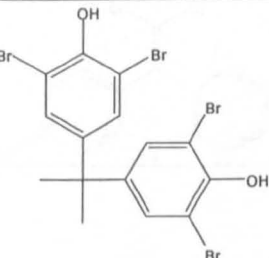
3-301		122-14-5	Fenitrothion	I	I	Test
3-303		140-56-7	Fenaminosulf	I	I	Train
3-306		6164-98-3	Chlordimeform	I	I	Val
3-307		21435-27-8	SER-TYR	I	I	Val
3-308		17138-28-2	Ethyl 4-hydroxyphenyl acetate	I	I	Train
3-309		370-14-9	p-Hydroxymethamphetamine	I	I	Train
3-314		1844-00-4	1,1-bis(4-Hydroxyphenyl)-isobutane	A	A	Train
4-001		92-52-4	Biphenyl	I	I	Val
4-002		2051-62-9	4-Chloro-1,1'-biphenyl	A	I	Val
4-003		2051-60-7	2-Chlorobiphenyl (PCB 1)	I	I	Train
4-004		2051-61-8	3-Chlorobiphenyl (PCB 2)	I	I	Train
4-005		92-86-4	p,p'-Dibromobiphenyl	I	I	Train
4-006		398-23-2	1,1'-Biphenyl, 4,4'-difluoro-	I	I	Train

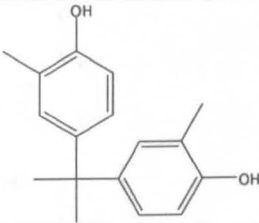
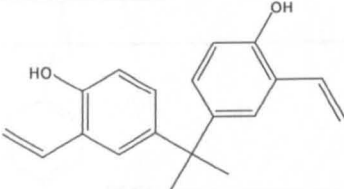
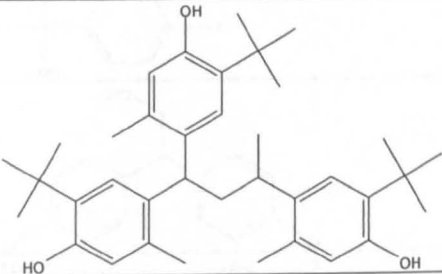
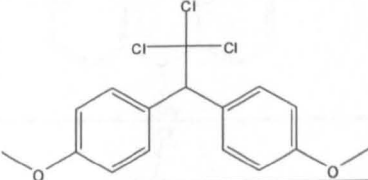
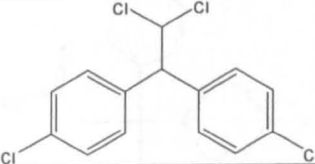
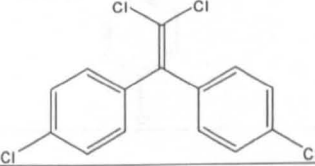
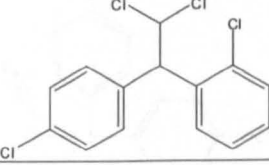
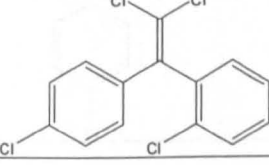
4-007		3001-15-8	4,4'- Diiodobiphenyl	A	I	Train
4-008		92-69-3	p- Phenylphenol	A	A	Train
4-009		580-51-8	m- Phenylphenol	I	A	Val
4-010		90-43-7	ortho- Phenylphenol	I	I	Test
4-011		92-88-6	4,4- Dihydroxydiph enyl	A	A	Test
4-012		1806-29-7	2,2'- Dihydroxybiph enyl = 2,2'- Biphenol	I	I	Train
4-013		491-45-2	Phloroglucide	I	I	Train
4-014		28034-99-3	4-Hydroxy-4'- chlorobiphenyl	A	A	Train
4-015		29558-77-8	4-(4- Bromophenyl) phenol	A	A	Train
4-016		16881-71-3	4'-Methoxy- biphenyl-4-ol	A	A	Test
4-017		58574-03-1	4'-Hydroxy-4- biphenylcarbo xylic acid	A	I	Train
4-018		19812-93-2	4'-Hydroxy-4- biphenylcarbo nitrile	A	A	Val
4-019		2417-04-1	3,3',5,5'- Tetramethyl- (1,1'-biphenyl)- 4,4'-diol	A	I	Val
4-020		613-37-6	4- Methoxybiphe nyl	I	I	Val
4-021		92-92-2	1,1'-Biphenyl- 4-carboxylic acid	I	A	Test

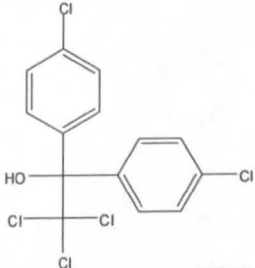
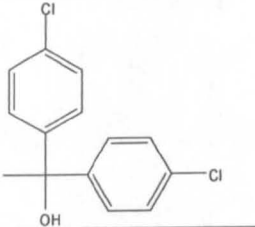
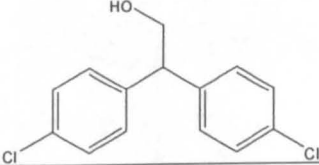
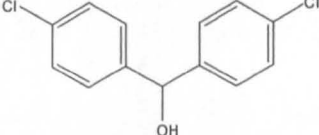
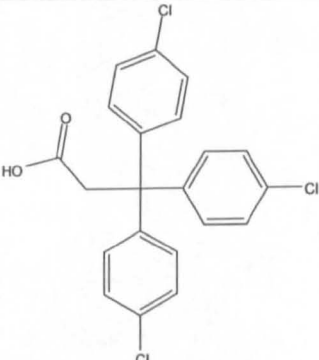
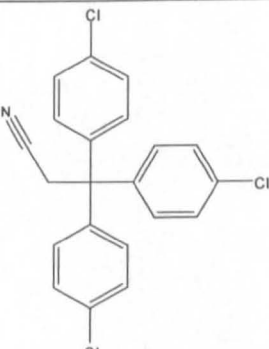
4-022		101-81-5	Diphenylmethane	I	I	Train
4-023		101-53-1	4-(Phenylmethyl)-phenol	A	A	Train
4-024		620-92-8	4,4'-Dihydroxydiphenylmethane	A	A	Train
4-025		2081-08-5	4,4'-Ethylidenebisphenol	A	A	Train
4-026		599-64-4	4-alpha-Cumylphenol	A	A	Train
4-027		80-05-7	Bisphenol A	A	A	Val
4-028		77-40-7	2,2-bis(4-Hydroxyphenyl)-butane	A	A	Train
4-029		6807-17-6	2,2-bis(4-Hydroxyphenyl)-4-methyl-n-pentane	A	A	Test
4-030		1571-75-1	4,4'-(1-Phenylethylidene)bisphenol	A	A	Test

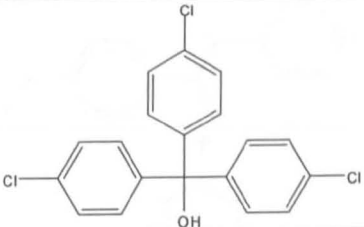
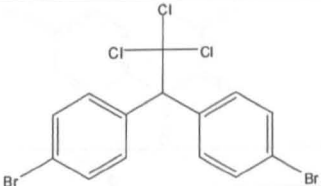
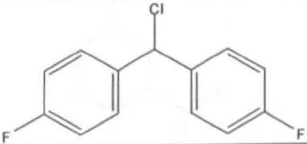
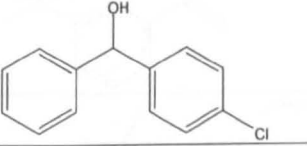

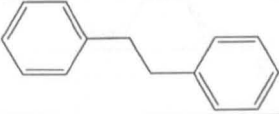
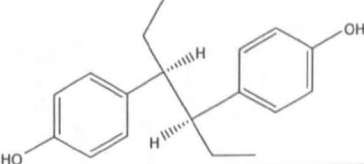
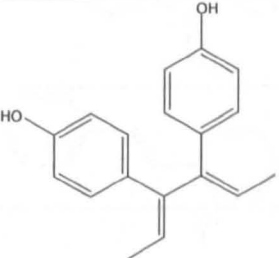
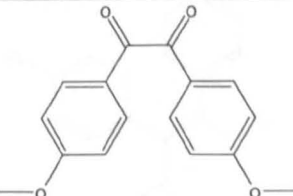
4-031		27955-94-8	tris(4-Hydroxyphenyl)ethane	A	A	Val
4-032		81-92-5	2-[bis(4-Hydroxyphenyl)methyl]benzyl alcohol = Phenolphthalein	A	A	Train
4-033		1478-61-1	4,4'-(Hexafluoroisopropylidene)diphenol	A	A	Train
4-034		843-55-0	4,4'-(Cyclohexylidene)bisphenol	A	A	Val
4-035		1943-97-1	4,4'-(Octahydro-4,7-methano-5H-inden-5-ylidene)bisphenol	A	I	Train
4-036		3236-71-3	4,4'-(9-Fluorenylidene)diphenol	A	I	Val

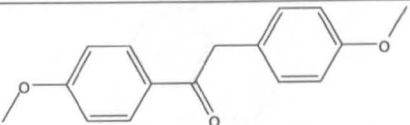
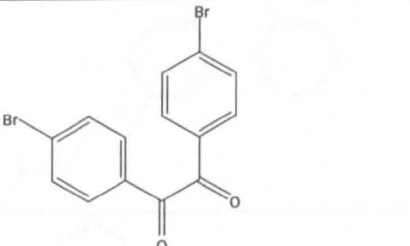
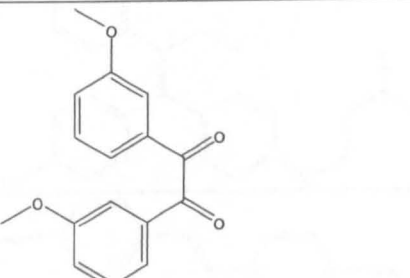
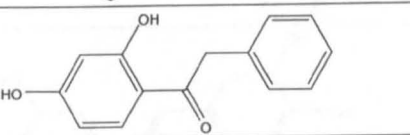
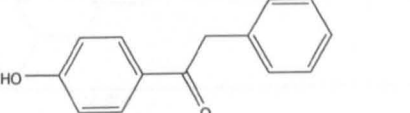
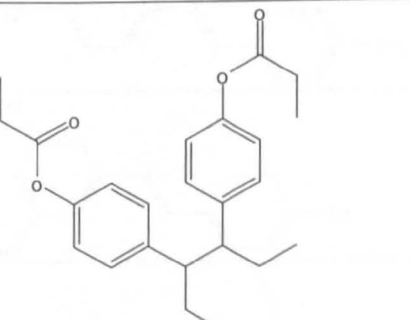

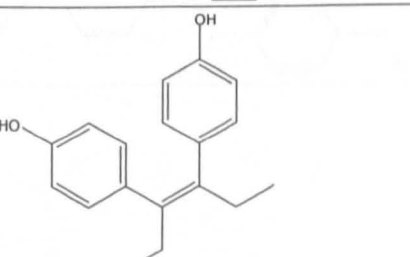
4-038		603-45-2	Rosolic acid	A	I	Train
4-039		143-74-8	Phenol red	I	I	Train
4-040		126-00-1	4,4-bis(4-Hydroxyphenyl) valeric acid	A	I	Train
4-041		13595-25-0	4,4'-(1,3-Phenylenediisopropylidene) bisphenol	A	I	Train
4-042		2167-51-3	4,4'-(1,4-Phenylenediisopropylidene) bisphenol	A	I	Test
4-043		37693-01-9	Clofoctol	A	I	Test

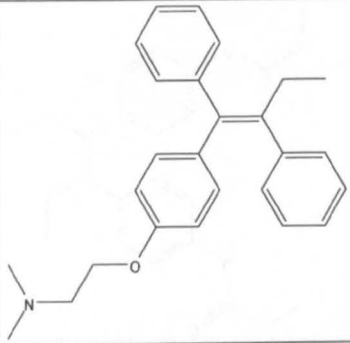
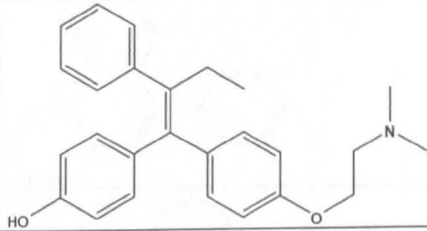
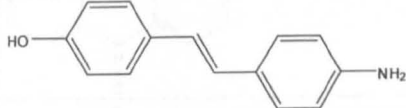
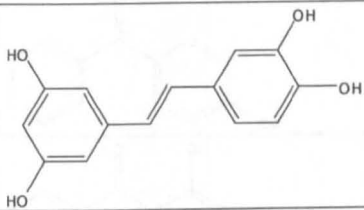
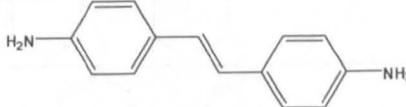
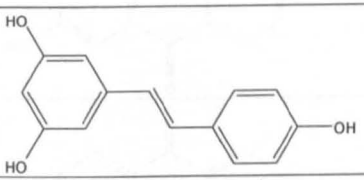

4-044		80-09-1	4,4'-Sulfonyldiphenol	A	A	Train
4-045		2784-27-2	5-phenyl-5-(p-hydroxy)phenylhydantoin	I	I	Train
4-047		2303-01-7	m-Cresol purple	A	A	Train
4-048		1844-01-5	4,4'-Dihydroxytetraphenylmethane	A	I	Train
4-049		5384-21-4	4,4'-Methylenebis(2,6-dimethylphenol)	A	I	Train
4-050		5613-46-7	2,2-bis-(3,5-Dimethyl-4-hydroxyphenyl)-propane	A	A	Test
4-051		79-94-7	Tetrabromobisphenol A (TBBP-A)	I	I	Train

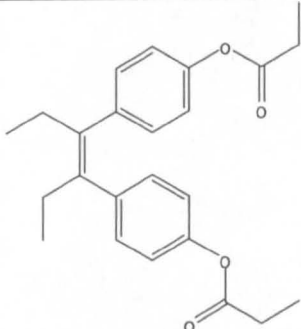

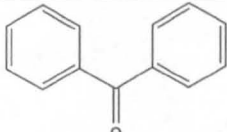
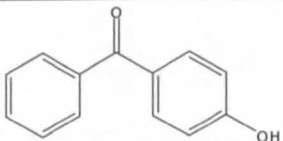
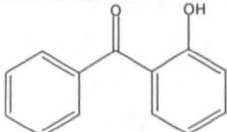
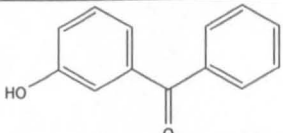
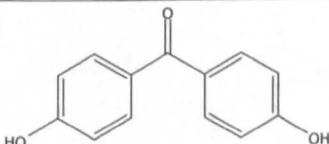
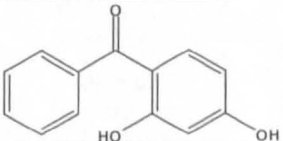
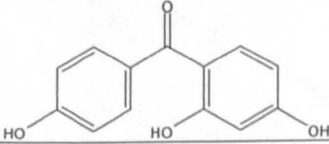
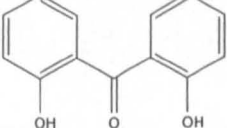
4-052		79-97-0	2,2-bis(4-Hydroxy-3-methylphenyl)propane	A	A	Train
4-053		1745-89-7	2,2'-Diallylbisphenol A	A	I	Test
4-054		1843-03-4	1,1,3-tris(2-Methyl-4-hydroxy-5-tertbutylphenyl)butane	A	I	Train
4-060		72-43-5	Methoxychlor	A	I	Val
4-063		72-54-8	Dichlorodiphenyldichloroethane	I	I	Train
4-064		72-55-9	p,p'-DDE	I	I	Test
4-065		53-19-0	o,p'-DDD	A	A	Train
4-066		3424-82-6	o,p'-DDE	A	I	Train

4-067		115-32-2	Dicofof = Kelthane	A	I	Train
4-068		80-06-8	Chlorfenethol	I	I	Train
4-069		2642-82-2	2,2-Bis(p- chlorophenyl)e thanol	I	A	Test
4-070		90-97-1	4,4'- Dichlorobenzh ydrol	I	I	Train
4-071		2168-06-1	3,3,3-tris (4- Chlorophenyl) propionic acid	I	I	Val
4-072		2172-51-2	3,3,3-tris (4- Chlorophenyl) propionitrile	I	I	Train

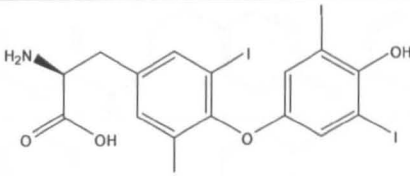
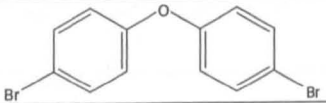
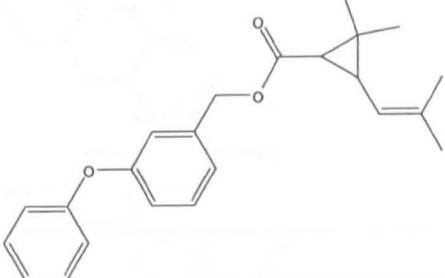
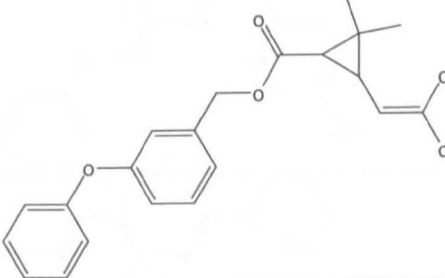
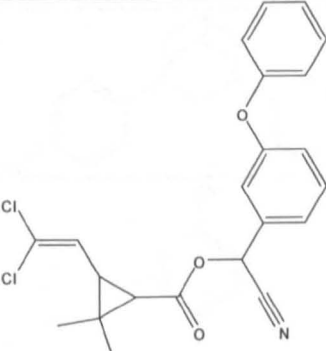
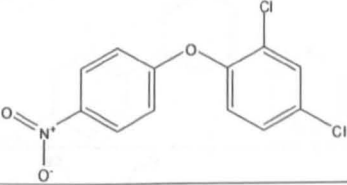
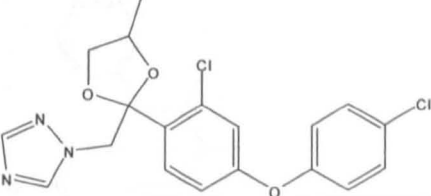
4-073		3010-80-8	4,4',4''-Trichlorotriphenyl alcohol	I	I	Test
4-074		2990-17-2	p,p'-Dibromodiphenyl trichloroethane	I	I	Train
4-076		27064-94-4	Chloro bis-(4-fluorophenyl)methane	I	A	Train
4-077		119-56-2	4-chloro-α-phenylbenzenemethanol	I	I	Test
4-078		22856-62-8	4,4'-Ethylenedi-m-toluidine	A	A	Train
4-080		103-29-7	Bibenzyl	I	I	Test
4-081		84-16-2	Hexestrol	A	A	Train
4-082		84-17-3	Dehydrostilbestrol	A	A	Train
4-083		1226-42-2	p-Anisil	I	I	Test

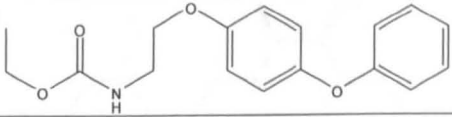
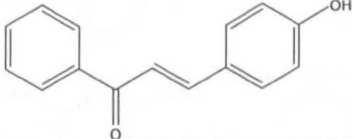
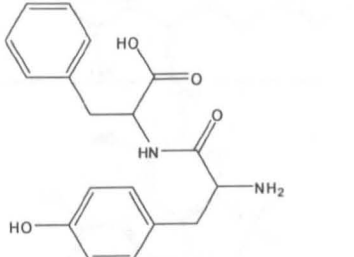
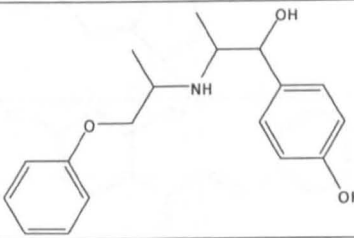
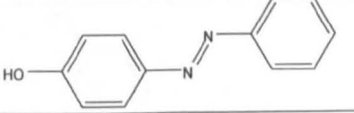
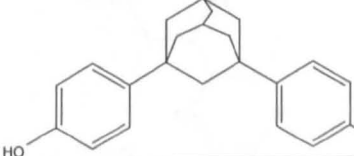
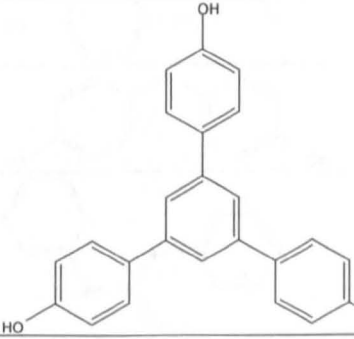
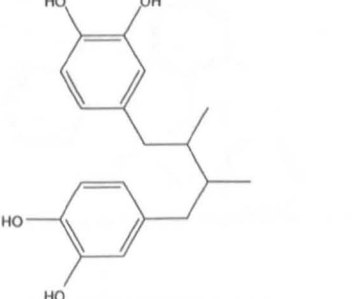
4-084		120-44-5	Desoxyanisoin	I	I	Train
4-085		35578-47-3	4,4'-Dibromobenzil	I	I	Train
4-086		40101-17-5	3,3'-Dimethoxybenzil	I	I	Train
4-087		3669-41-8	Benzyl 2,4-dihydroxyphenyl ketone	A	A	Train
4-088		2491-32-9	Benzyl 4-hydroxyphenyl ketone	A	A	Test
4-089		4825-53-0	Hexestrol dipropionate	A	A	Train
4-090		103-30-0	trans-Stilbene	I	I	Test
4-091		56-53-1	diethylstilbestrol	A	A	Train

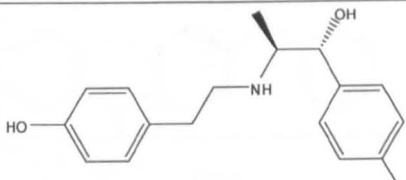
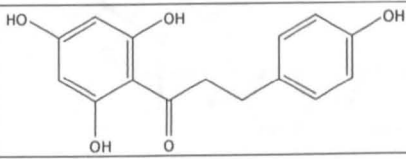
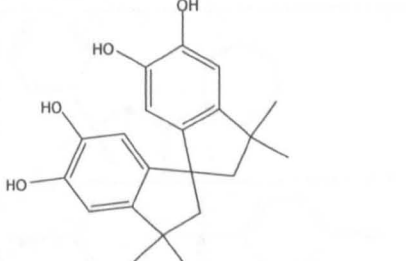
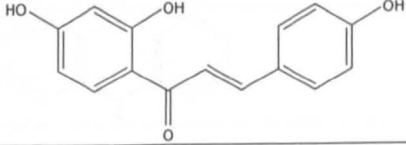
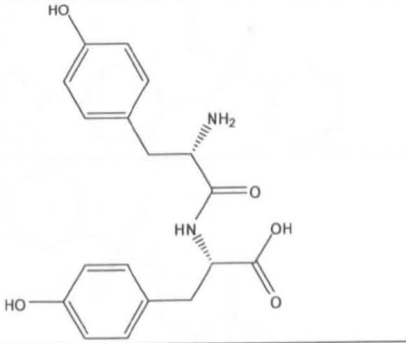
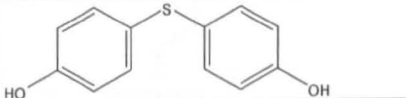
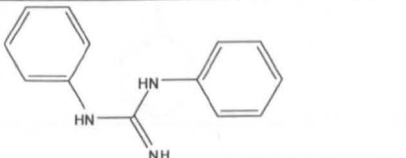
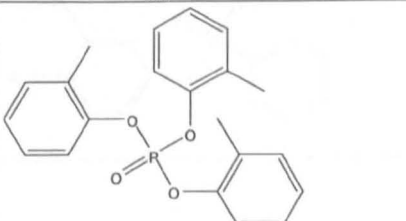
4-092		10540-29-1	Tamoxifen	A	I	Train
4-093		68047-06-3	4-Hydroxytamoxifen	A	I	Val
4-094		836-44-2	4-Amino-4'-hydroxystilbene	A	A	Train
4-095		10083-24-6	Piceatannol	A	A	Test
4-096		54760-75-7	4,4'-Diaminostilbene dihydrochloride	I	I	Train
4-097		501-36-0	Resveratrol	A	A	Train
4-099		569-57-3	Chlorotrianisene	A	I	Test

4-100		130-80-3	Diethylstilbestrol dipropionate	A	A	Val
4-101		58-72-0	Triphenylethylene	A	A	Train
4-102		119-61-9	Benzophenone	I	I	Train
4-103		1137-42-4	4-Hydroxybenzophenone	A	A	Test
4-104		117-99-7	2-Hydroxybenzophenone	I	I	Val
4-105		13020-57-0	3-Hydroxybenzophenone	A	A	Train
4-106		611-99-4	4,4'-Dihydroxybenzophenone	A	A	Train
4-107		131-56-6	2,4-Dihydroxybenzophenone	A	A	Val
4-108		1470-79-7	2,4,4'-Trihydroxybenzophenone	A	A	Val
4-109		835-11-0	2,2'-Dihydroxybenzophenone	I	I	Train

4-110		31127-54-5	2,3,4,4'- Tetrahydroxyb enzophenon	A	I	Train
4-111		90-96-0	4,4'- Dimethoxyben zophenone	I	I	Val
4-112		131-55-5	2,2',4,4'- Tetrahydroxyb enzophenone	A	A	Test
4-113		42019-78-3	4-Chloro-4'- hydroxybenzo phenone	A	A	Train
4-114		25913-05-7	4-Fluoro-4'- hydroxybenzo phenone	A	A	Train
4-115		1143-72-2	2,3,4- Trihydroxyben zophenone	A	I	Test
4-116		90-98-2	4,4'- Dichlorobenzo phenone	I	I	Val
4-117		3988-03-2	4,4'- Dibromobenzo phenone	A	A	Train
4-118		831-82-3	4- Phenoxypheno l	A	A	Train
4-119		103-16-2	Hydroquinone monobenzylet her	A	A	Test
4-120		1965-09-9	4,4'- Oxydiphenol	A	A	Val
4-121		7005-72-3	1-Chloro-4- phenoxybenze ne	I	I	Train

4-122		51-48-9	L-Thyroxine	I	I	Val
4-123		2050-47-7	p,p'-Dibromodiphenyl ether	I	I	Train
4-124		26002-80-2	Fenothrin	I	I	Train
4-125		52645-53-1	Permethrin	I	I	Test
4-127		52315-07-8	Cypermethrin	I	I	Val
4-128		1836-75-5	Nitrofen	I	I	Train
4-129		119446-68-3	Difenconazole	I	I	Val

4-130		79127-80-3	Fenoxycarb	I	I	Train
4-132		20426-12-4	4-Hydroxychalcone	A	A	Val
4-134		17355-11-2	TYR-PHE	I	A	Train
4-135		579-56-6	Isoxsuprine hydrochloride	I	I	Train
4-137		1689-82-3	4-Hydroxyazobenzene	A	A	Train
4-139		37677-93-3	4,4'-(1,3-Adamantanediyl)diphenol	A	A	Train
4-140		15797-52-1	1,3,5-Tris(4-hydroxyphenyl)benzene	A	I	Train
4-141		500-38-9	Nordihydroguaiaretic acid	A	I	Train

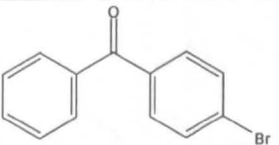
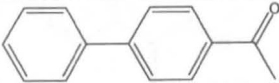
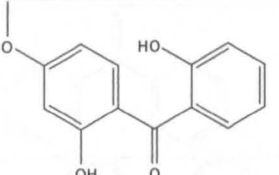
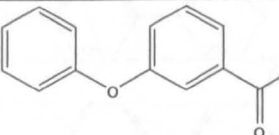
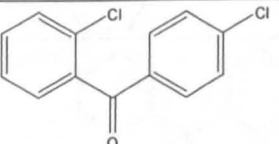
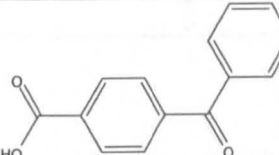
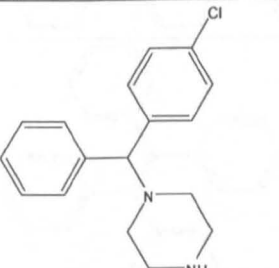
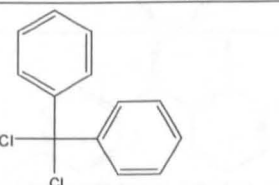
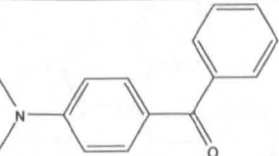
4-142		23239-51-2	Ritodrine hydrochloride	I	I	Train
4-143		60-82-2	Phloretin	A	A	Train
4-144		77-08-7	3,3,3',3'-Tetramethyl-1,1'-spirobisindane-5,5',6,6'-tetrol	A	A	Train
4-145		961-29-5	Isoliquirtigenin	A	A	Val
4-146		1050-28-8	TYR-TYR	I	I	Train
4-147		2664-63-3	4,4'-Thiobisphenol	A	A	Train
4-148		102-06-7	Diphenylguanidine	I	I	Train
4-149		78-30-8	Tris(o-cresyl) phosphate	A	I	Train

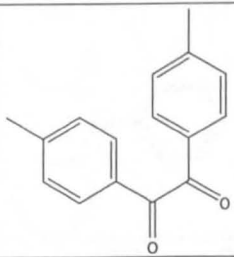
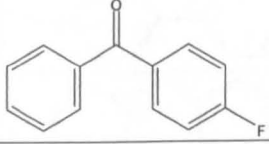
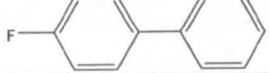
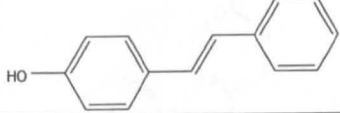
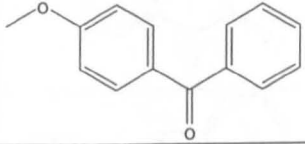
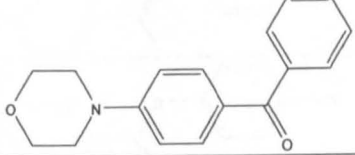

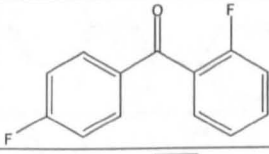
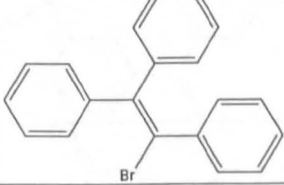
4-150		74-31-7	Diphenyl-p-phenylenediamine	A	I	Train
4-152		65277-42-1	Ketoconazole	I	I	Train
4-153		61432-55-1	S-(1-methyl-1-phenylethyl)pyrrolidine-1-carbothioate	I	I	Train
4-154		80-08-0	4,4'-Sulfonylbisbenzenamine	I	I	Val
4-156		122-39-4	Diphenylamine	I	I	Test
4-157		2104-64-5	O-Ethyl O-4-nitrophenylphenylphosphonothioate ; EPN	I	I	Train
4-158		17606-31-4	Bensultap; 1,3-bis(phenylsulfonylthio)-2-(N,Ndimethylamino)propane	A	I	Val
4-159		21609-90-5	Leptophos (ISO) ; O-4-bromo-2,5-dichlorophenyl O-methylphenylphosphorothioate	A	I	Train

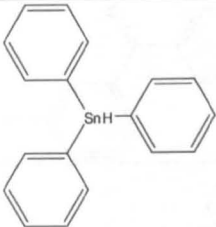
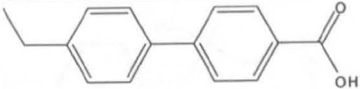
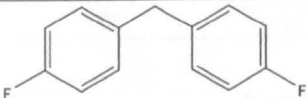
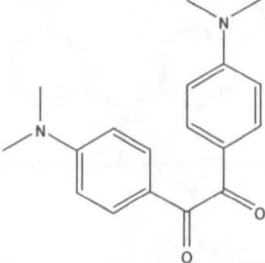
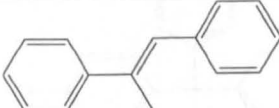
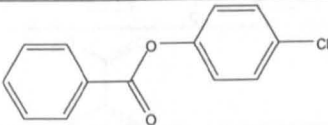
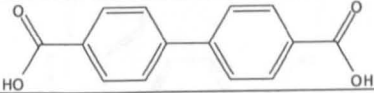
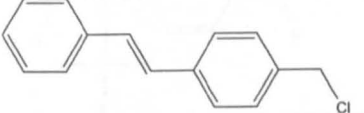
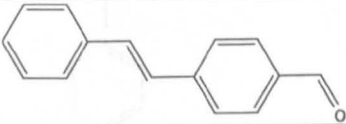
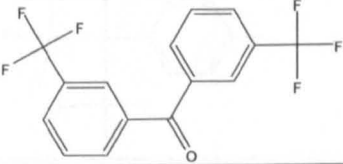
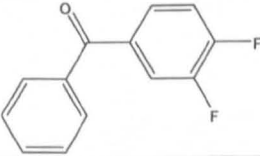
4-160		33089-61-1	Amitraz	I	I	Train
4-161		34123-59-6	Isoproturon	I	A	Val
4-162		142172-97-2	2-(4-Hydroxyphenyl)-5-pyrimidinol	I	I	Train
4-163		2051-85-6	Sudan orange G	A	A	Train
4-164		3294-03-9	2,2'-Thiobis[4-(1,1,3,3-Tetramethylbutyl)phenol]	A	I	Train
4-165		3483-82-7	N-Benzoyl-L-tyrosine ethylester	I	I	Test
4-166		49745-95-1	Dobutamine hydrochloride	A	I	Train
4-167		74-39-5	Azo violet	I	I	Test
4-169		51-24-1	3,3'-Trilodothyroacetic acid	A	I	Test

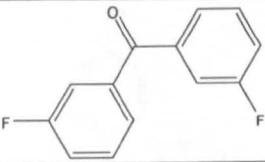
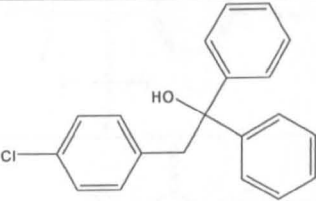
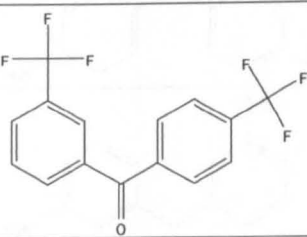
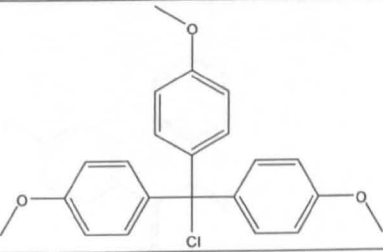
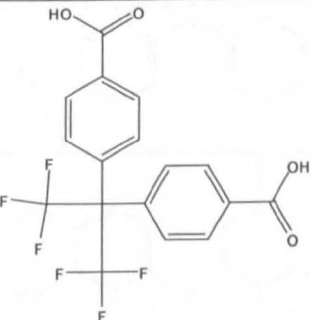
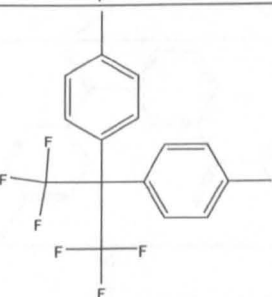
4-170		6893-02-3	3,5,3'-Triiodo-L-thyronine	I	I	Train
4-171		5817-39-0	L-3,3',5'-Triiodothyronine	I	I	Test
4-172		50-41-9	Clomiphene citrate (cis and trans mixture)	A	I	Train
4-174		35367-38-5	Diflubenzuron	I	I	Train
4-177		76-87-9	Fentin hydroxide	A	I	Test
4-179		2894-87-3	2,3,5,6-Tetrafluoro-4-(pentafluorophenyl)phenol	A	I	Val
4-182		728-87-0	4,4'-Dimethoxybenzhydrol	I	I	Train
4-189		101-77-9	4,4'-Methylenebisbenzeneamine	I	I	Train
4-190		101-80-4	4,4'-oxybisbenzenamine	I	I	Val

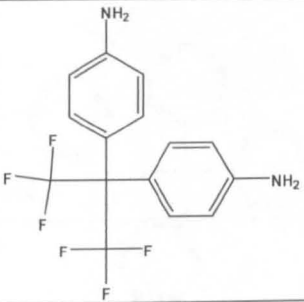
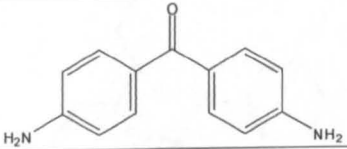
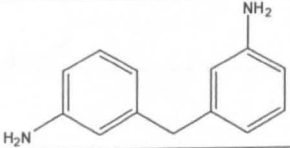
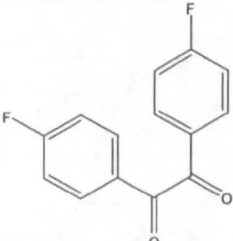
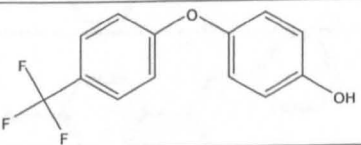
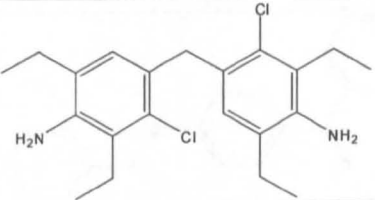
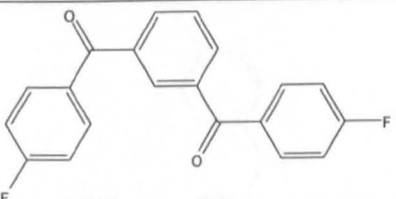
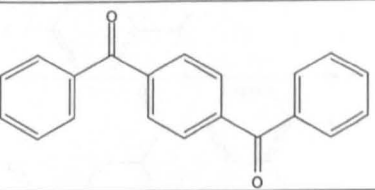
4-191		22494-42-4	Diflunisal	I	I	Train
4-192		92-66-0	4-Bromobiphenyl	I	I	Test
4-193		1144-74-7	4-Nitrobenzophenone	I	I	Train
4-195		118-55-8	Phenyl salicylate	I	I	Train
4-197		1620-21-9	Chlorcyclizine hydrochloride	I	I	Train
4-198		980-71-2	Brompheniramine	I	I	Train
4-199		434-90-2	Decafluorobiphenyl	I	I	Train
4-200		3218-36-8	(1,1'-Biphenyl)-4-carboxaldehyde	I	I	Test
4-201		1016-78-0	m-Chlorobenzophenone	I	I	Val
4-202		134-85-0	p-Chlorobenzophenone	I	I	Train

4-203		90-90-4	p-Bromobenzophenone	I	I	Test
4-204		92-91-1	p-Acetylbiphenyl	I	I	Train
4-205		131-53-3	2,2'-Dihydroxy-4-methoxybenzophenone	I	I	Train
4-206		3739-38-6	3-Phenoxybenzoic acid	I	I	Train
4-207		85-29-0	Methanone, (2-chlorophenyl)(4-chlorophenyl)-	I	I	Train
4-210		611-95-0	4-Benzoylbenzoic acid	I	I	Test
4-211		303-26-4	1-(4-Chlorobenzhydryl)piperazine	I	I	Train
4-212		2051-90-3	Dichlorodiphenylmethane	I	I	Val
4-213		530-44-9	4-(Dimethylamino)benzophenone	I	I	Test

4-214		3457-48-5	4,4'-Dimethylbenzil	I	I	Train
4-215		345-83-5	4-Fluorobenzophenone	I	I	Val
4-216		324-74-3	4-Fluorobiphenyl	I	I	Val
4-217		6554-98-9	trans-4-Hydroxystilbene	A	A	Train
4-218		611-94-9	4-Methoxybenzophenone	I	I	Val
4-219		24758-49-4	4-Morpholinobenzophenone	I	I	Train
4-220		632-51-9	Tetraphenylethylene	A	I	Train
4-221		342-25-6	2,4'-Difluorobenzophenone	I	I	Train
4-222		1607-57-4	Bromotriphenylethylene	A	A	Train


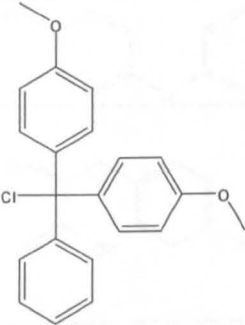
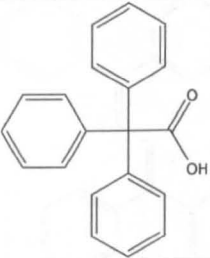
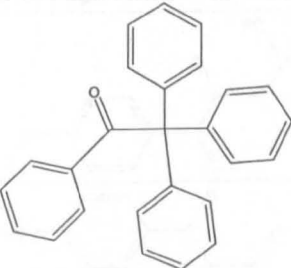
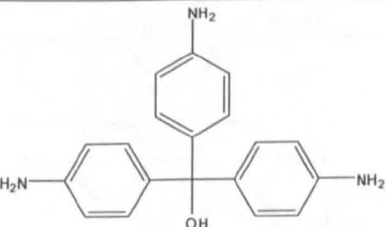
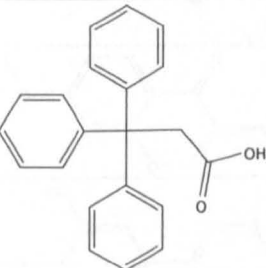
4-223		892-20-6	Triphenyltin hydride	A	I	Val
4-224		5731-13-5	4-Ethylbiphenyl-4'-carboxylic acid	I	I	Train
4-225		457-68-1	4,4'-Difluorodiphenylmethane	I	I	Train
4-226		17078-27-2	4,4'-Bis(dimethylamino)benzil	I	I	Train
4-227		833-81-8	Stilbene, alpha-methyl-, (E)-	I	I	Train
4-229		2005-08-5	4-Chlorophenyl benzoate	I	I	Val
4-230		787-70-2	4,4'-Biphenyldicarboxylic acid	I	I	Test
4-231		150253-59-1	4-Chloromethylstilbene	I	A	Train
4-232		40200-69-9	trans-4-Stilbenecarboxaldehyde	I	I	Train
4-233		1868-00-4	3,3'-Bis(trifluoromethyl)benzophenone	I	I	Test
4-234		85118-07-6	3,4-Difluorobenzophenone	I	I	Test

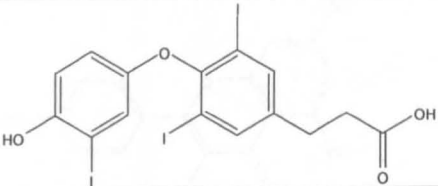
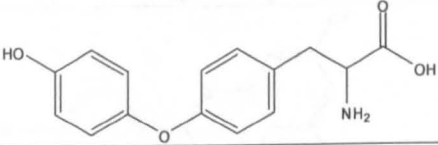
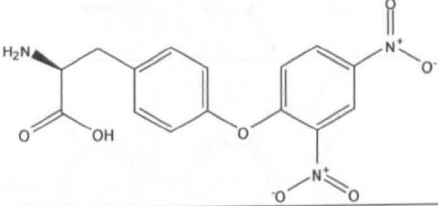
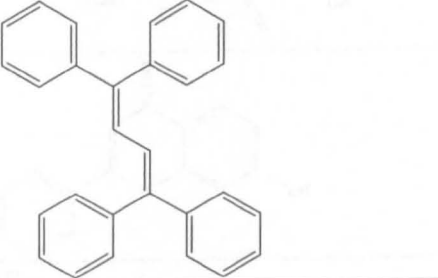
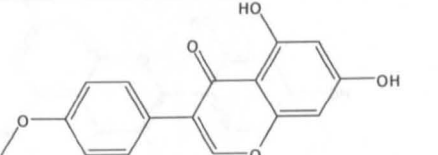
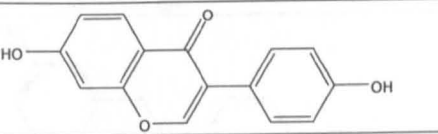
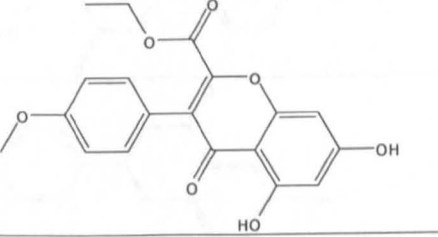
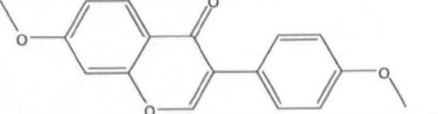
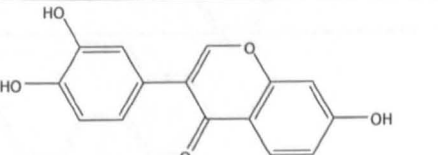
4-235		345-70-0	3,3'- Difluorobenzo- phenone	I	I	Train
4-236		109936-21-2	2-(4- Chlorophenyl)- 1,1- diphenylethan- ol	A	I	Val
4-237		21084-22-0	3,4'- Bis(trifluorome- thyl)benzophe- none	I	I	Val
4-238		49757-42-8	4,4',4''- Trimethoxytrity- l chloride	A	I	Train
4-239		1171-47-7	4,4'- (Hexafluoroiso- propylidene)bi- s(benzoic acid)	A	I	Train
4-240		1095-77-8	4,4'- (Hexafluoroiso- propylidene)di- toluene	A	I	Test

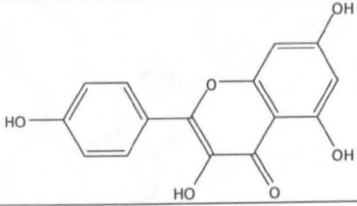
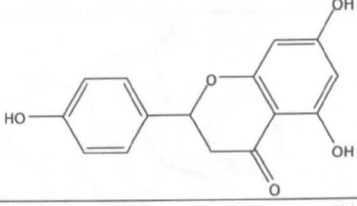
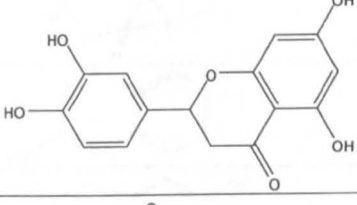
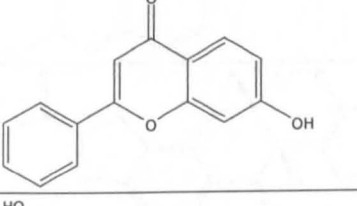
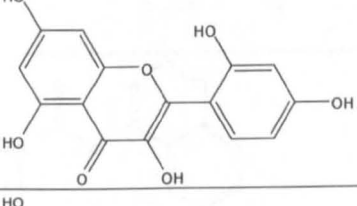
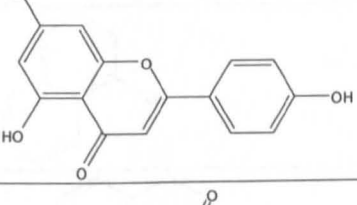

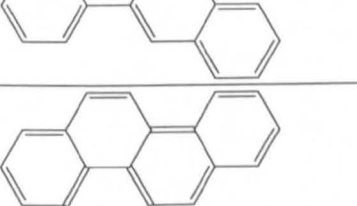
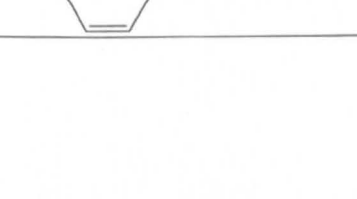
4-241		1095-78-9	4,4'-(Hexafluoroisopropylidene)di-aniline	A	A	Val
4-242		611-98-3	4,4'-Diaminobenzophenone	I	I	Train
4-243		19471-12-6	3,3'-Methylenedianiline	I	I	Train
4-244		579-39-5	4,4'-Difluorobenzil	I	I	Train
4-245		39634-42-9	4-(4-(Trifluoromethyl)phenoxy)phenol	A	I	Train
4-247		106246-33-7	4,4'-Methylenebis(3-chloro-2,6-diethylaniline)	A	I	Train
4-249		108464-88-6	1,3-Bis(4-fluorobenzoyl)benzene	I	I	Train
4-250		3016-97-5	1,4-Dibenzoylbenzene	I	I	Train

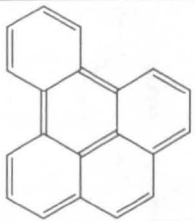
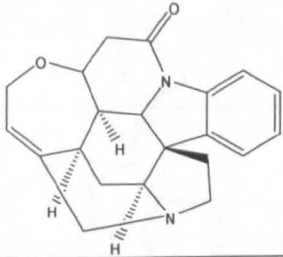
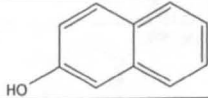
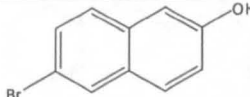
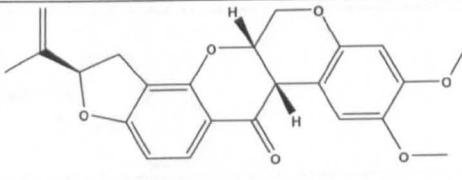
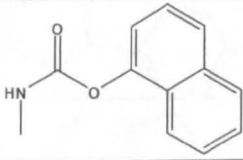
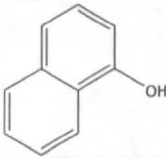
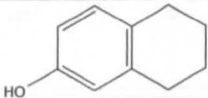
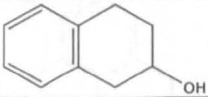
4-251	 <chem>CN(C)C1=CC=C(C=C1)C2(C)C3=CC=C(N)C=C3C2(C)C4=CC=C(N)C=C4</chem>	2687-27-6	4,4'-(1,3-Phenylenediisopropylidene)bis(4-aminobenzene)	A	I	Train
4-252	 <chem>O=CC1=CC(OC2=CC=CC=C2)=C(OC3=CC=CC=C3)C=C1</chem>	5447-02-9	3,4-Dibenzylbenzaldehyde	A	I	Val
4-253	 <chem>CC(C)OC(=O)C(C)(C)OC(=O)C1=CC=C(C=C1)C(=O)C2=CC=C(Cl)C=C2</chem>	49562-28-9	Fenofibrate	I	I	Test
4-256	 <chem>ClC1=CC=C(C=C1)C(=O)CC2=CC=CC=C2</chem>	1889-71-0	Benzyl 4-chlorophenyl ketone	I	I	Train
4-257	 <chem>BrC1=CC=C(C=C1)C(=O)CC2=CC=CC=C2</chem>	2001-29-8	Benzyl 4-bromophenyl ketone	I	I	Val
4-258	 <chem>O=C(C1=CC=CC=C1)C2=CC=C(OC3=CC=CC=C3)C=C2</chem>	54589-41-2	4-Benzyloxybenzophenone	I	I	Train
4-259	 <chem>Fc1ccc(cc1)C(=O)/C=C/C(=O)c2ccc(F)cc2</chem>	25650-13-9	trans-1,2-Bis(4-fluorobenzoyl)ethylene	A	A	Train
4-260	 <chem>Cc1ccc(cc1C)C(=O)c2ccc([N+](=O)[O-])cc2</chem>	42187-33-7	3,4-Dimethyl-3'-nitrobenzophenone	I	I	Train

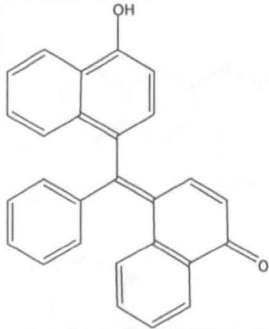
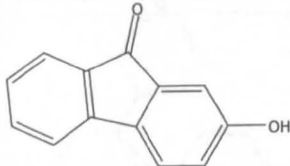
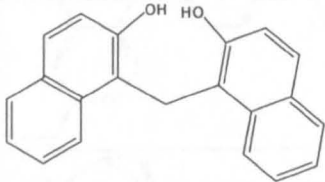
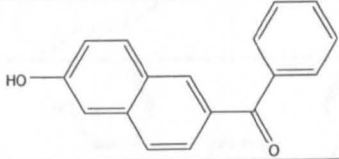
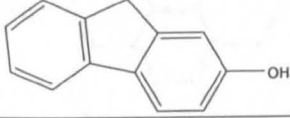
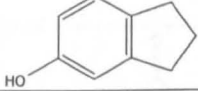
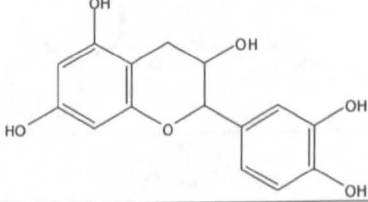
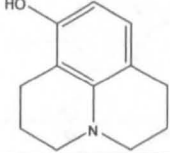
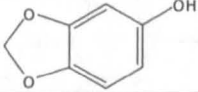
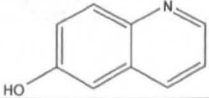
4-261		143130-82-9	Dimethyl cis-stilbene-4,4'-dicarboxylate	I	I	Train
4-264		133005-88-6	cis-Stilbene-4,4'-dicarboxylic acid	I	I	Val
4-265		101-61-1	4,4'-Methylene bis(N,N'-dimethylaniline)	I	I	Train
4-270		3539-42-2	Bisphenol A O,O-Diacetic acid	I	I	Train
4-271		2772-45-4	2,4-bis(alpha,alpha dimethylbenzyl)Phenol	A	I	Val
4-273		1675-54-3	2,2'-bis(4-(2,3-epoxypropoxy)phenyl)Propane	I	I	Train
4-276		596-27-0	o-Cresolphthalein	A	I	Train

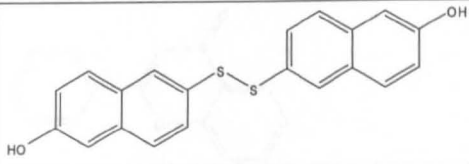
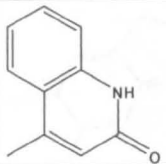
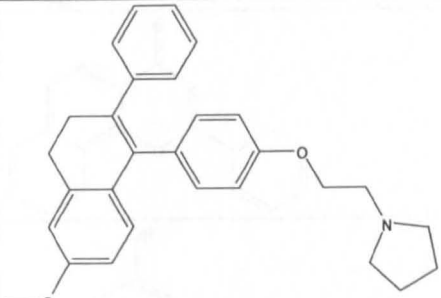
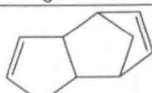
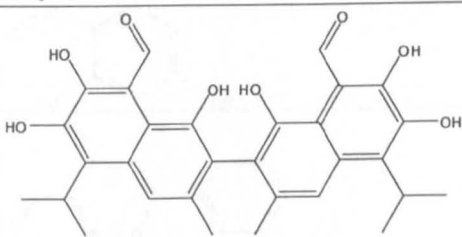
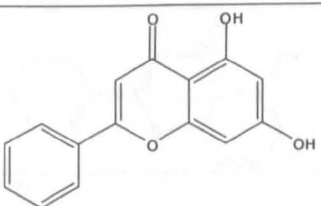
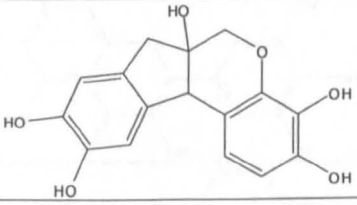
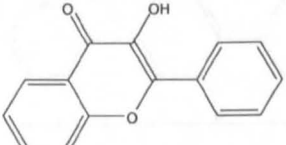
4-277		519-73-3	Triphenylmethane	A	I	Test
4-279		40615-36-9	4,4'-Dimethoxytrityl chloride	I	I	Train
4-281		595-91-5	Triphenylacetic acid	I	I	Val
4-282		466-37-5	2,2,2-Triphenylacetophenone	A	I	Train
4-284		467-62-9	Tris-(4-amino-phenyl)-methanol	A	I	Val
4-285		900-91-4	3,3,3-Triphenylpropionic acid	I	I	Test

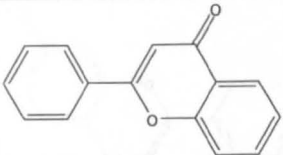
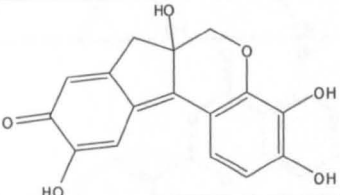
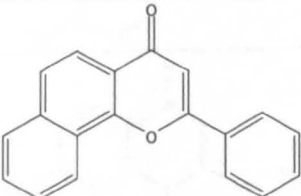
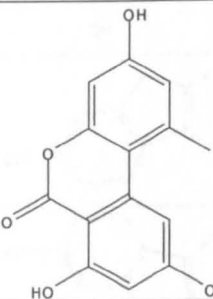
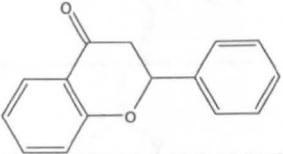
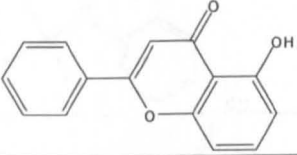
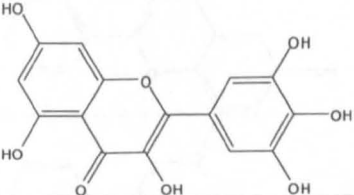
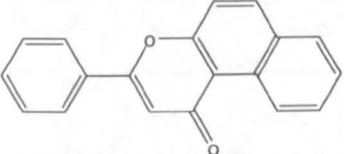
4-288		51-26-3	3,3',5-Triiodothyropropionic acid	A	I	Train
4-292		1596-67-4	L-Thyronine	I	I	Train
4-293		10567-73-4	O-Mono-2,4-DNP-L-tyrosine	I	I	Test
4-295		1450-63-1	1,1,4,4-Tetraphenyl-1,3-butadiene	A	I	Val
5-001		491-80-5	Biochanin A	A	A	Val
5-002		486-66-8	Daidzein	A	A	Train
5-003		15485-76-4	2-Carbethoxy-5,7-dihydroxy-4'-methoxyisoflavone	A	A	Train
5-004		1157-39-7	4',7-Dimethoxyisoflavone	I	I	Train
5-005		485-63-2	3',4',7-Trihydroxyisoflavone	A	A	Train

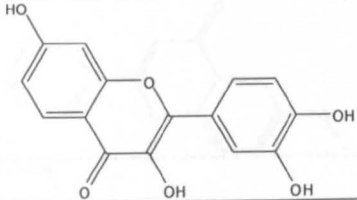
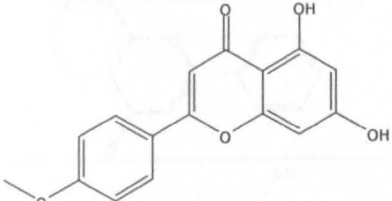
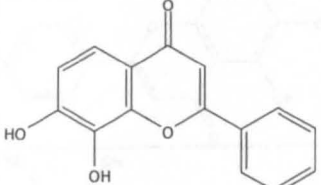
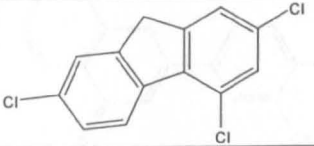
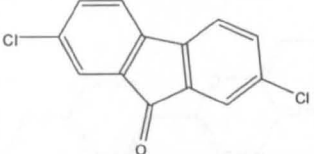
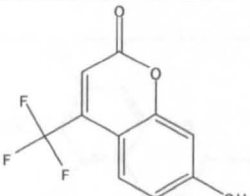
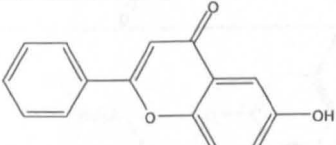
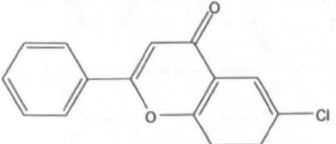
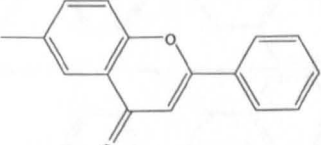
5-006		520-18-3	Kaempferol	A	A	Train
5-007		480-41-1	Naringenin	A	A	Train
5-008		491-70-3	Luteolin	A	A	Test
5-009		6665-86-7	7-Hydroxy-2-phenyl-4H-1-benzopyran-4-one	I	A	Val
5-011		480-16-0	Morin	A	A	Val
5-012		520-36-5	Apigenin	A	A	Train
5-014		4250-77-5	6-Hydroxyflavone	A	A	Train
5-016		56-55-3	Benz(a)anthracene	A	A	Train
5-017		50-32-8	Benzo[a]pyrene	I	I	Test

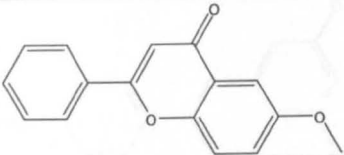
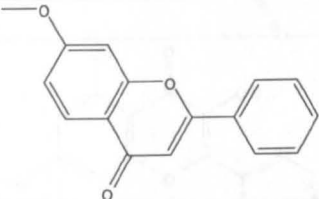
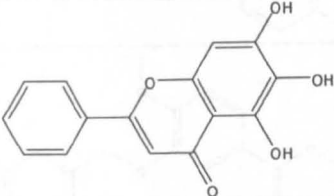
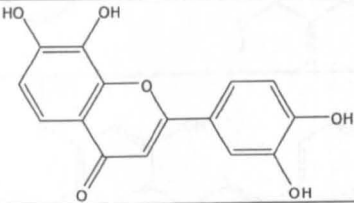
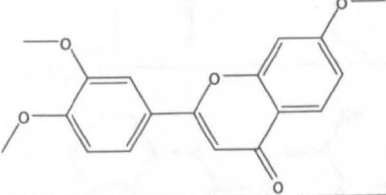
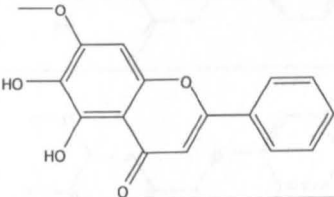
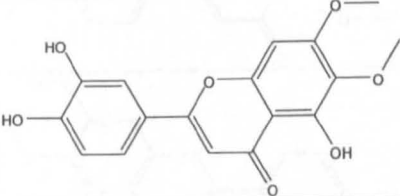
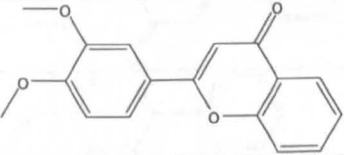
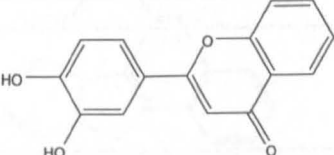
5-019		192-97-2	Benzo[e]pyrene	I	A	Train
5-020		57-24-9	Strychnine	I	I	Train
5-021		135-19-3	2-Naphthol	A	I	Test
5-022		15231-91-1	6-Bromo-2-naphthol	A	A	Train
5-023		83-79-4	Rotenone	I	I	Val
5-024		63-25-2	Carbaryl	I	I	Train
5-025		90-15-3	1-Naphthol	I	I	Train
5-026		1125-78-6	5,6,7,8-Tetrahydro-2-naphthol = 6-Hydroxytetralin	A	I	Train
5-027		530-91-6	Tetrahydronaphthol-2	I	I	Train

5-028		145-50-6	p-Naphtholbenzoin	A	A	Test
5-029		6949-73-1	2-Hydroxy-9-fluorenone	A	A	Train
5-030		1096-84-0	1,1'-Methylenedi-2-naphthol	A	I	Test
5-031		52222-87-4	6-Benzoyl-2-naphthol	A	A	Train
5-032		2443-58-5	2-Hydroxyfluorene	A	A	Val
5-034		1470-94-6	5-Hydroxyindan	I	I	Val
5-035		490-46-0	L-Epicatechin	A	I	Train
5-036		41175-50-2	8-Hydroxyjulolidine	I	I	Train
5-038		533-31-3	Sesamol	I	I	Train
5-039		580-16-5	6-Hydroxyquinoline	I	I	Val

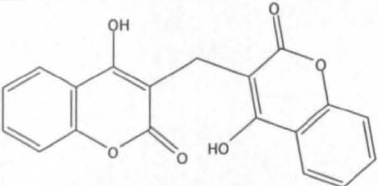
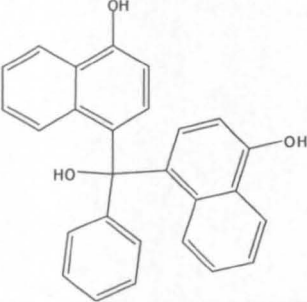
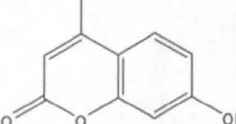
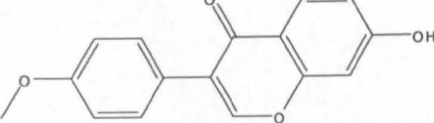
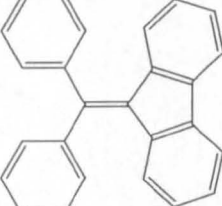
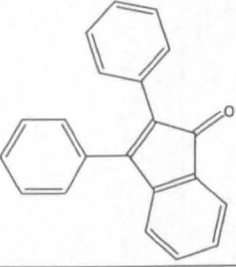
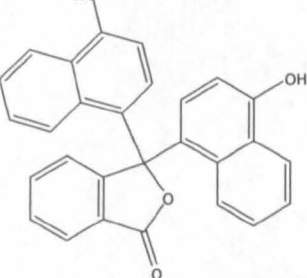
5-040		6088-51-3	6-Hydroxy-2-naphthyl disulfide	A	A	Train
5-042		607-66-9	2-Hydroxy-4-methylquinoline	I	I	Train
5-044		1847-63-8	Nafoxidine	A	I	Test
5-045		77-73-6	Dicyclopentadiene	I	I	Train
5-046		303-45-7	Gossypol	I	I	Train
5-047		480-40-0	5,7-Hydroxyflavone	I	A	Train
5-048		517-28-2	Hematoxylin	A	I	Train
5-049		577-85-5	3-Hydroxyflavone	I	I	Train

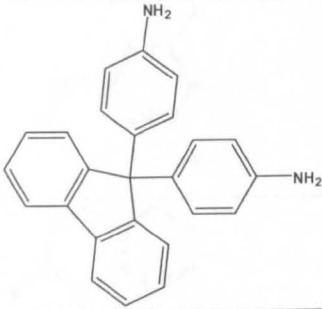
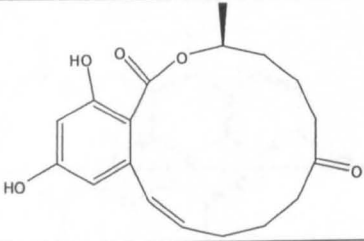
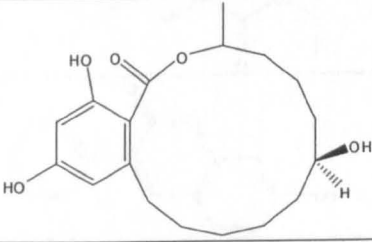
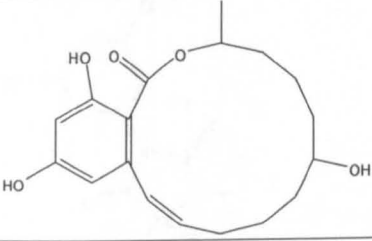
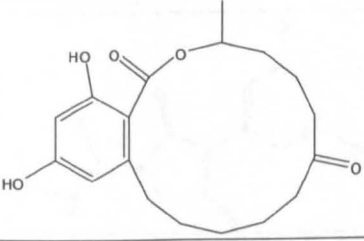
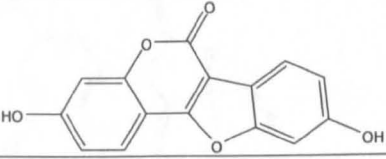
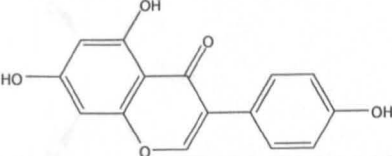
5-050		525-82-6	Flavone	I	I	Val
5-051		475-25-2	Hematein	A	I	Train
5-052		604-59-1	alpha-Naphthoflavone	I	I	Train
5-054		641-38-3	Alternariol	A	A	Train
5-055		487-26-3	Flavanone	I	I	Train
5-056		491-78-1	5-Hydroxyflavone	I	I	Train
5-057		529-44-2	Myricetin	A	I	Val
5-058		6051-87-2	5,6-Benzoflavone	I	I	Train

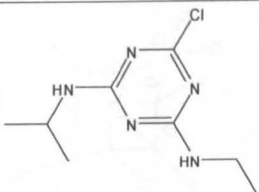
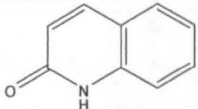
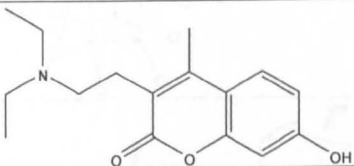
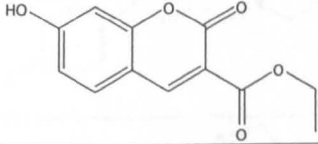
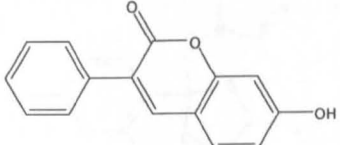
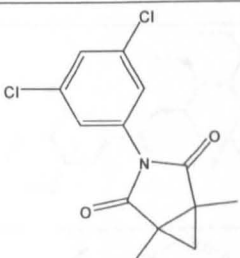
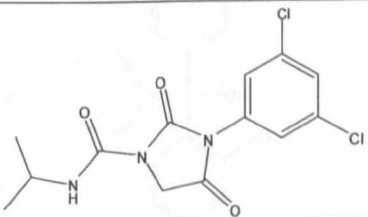
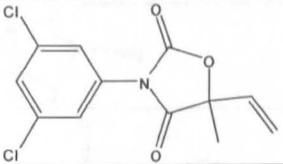
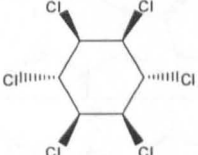
5-059		528-48-3	Fisetin	A	I	Train
5-060		480-44-4	Acacetin	I	A	Test
5-061		38183-03-8	7,8-Dihydroxy-2-phenyl-4h-1-benzopyran-4-one	A	I	Train
5-064		7061-81-6	2,4,7-Trichlorofluorene	I	I	Train
5-065		6297-11-6	2,7-Dichloro-9-fluorenone	I	I	Train
5-066		575-03-1	7-Hydroxy-4-(trifluoromethyl)coumarin	A	I	Train
5-067		6665-83-4	6-Hydroxyflavone	A	A	Train
5-068		10420-73-2	6-Chloroflavone	I	I	Train
5-069		29976-75-8	6-Methylflavone	I	A	Train

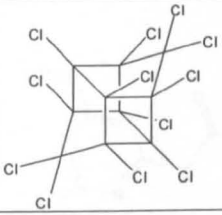
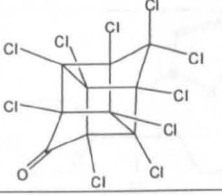
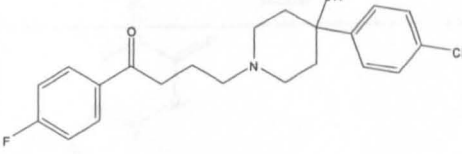
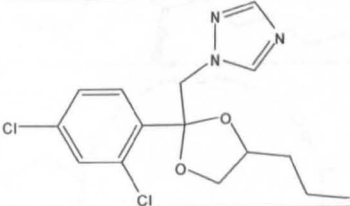
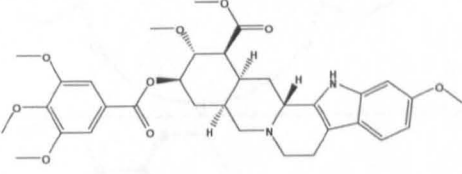
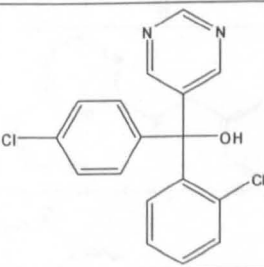
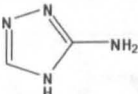
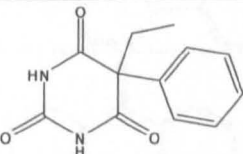
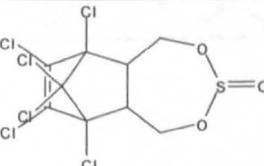
5-070		26964-24-9	6-Methoxyflavone	I	I	Test
5-071		22395-22-8	7-Methoxyflavone	I	I	Train
5-072		491-67-8	Baicalein	A	I	Train
5-075		3440-24-2	3',4',7,8-Tetrahydroxyflavone	A	I	Test
5-077		22395-24-0	3',4',7-Tromethoxyflavone	I	I	Train
5-078		29550-13-8	5,6-Dihydroxy-7-methoxyflavone	A	I	Val
5-079		34334-69-5	6,7-Dimethoxy-3',4',5-trihydroxyflavone	A	I	Train
5-080		4143-62-8	3',4'-Dimethoxyflavone	I	I	Train
5-081		4143-64-0	3',4'-Dihydroxyflavone	A	I	Test

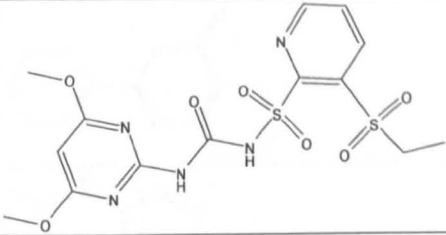
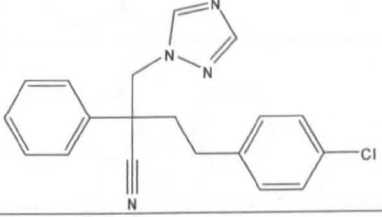
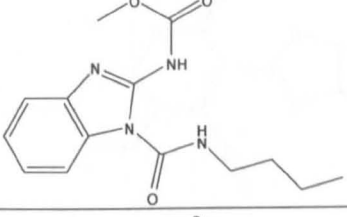
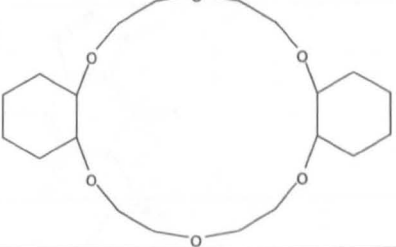
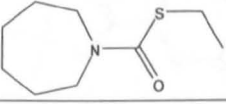
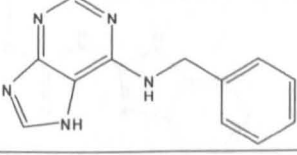
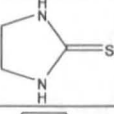
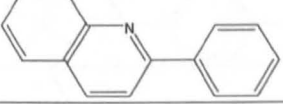
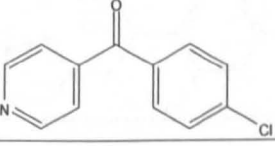
5-082		437-64-9	4',5-Dihydroxy-7-methoxyflavone	A	I	Train
5-083		491-54-3	4'-Methoxy-3,5,7-trihydroxyflavone	I	A	Val
5-084		62507-01-1	3'-Benzyloxy-5,7-dihydroxy-3,4'-dimethoxyflavone	A	I	Train
5-085		73694-15-2	2'-Hydroxy-2,4,4',5',6'-pentamethoxychalcone	I	I	Train
5-086		855-96-9	3',5-Dihydroxy-4',6,7-trimethoxyflavone	I	I	Test
5-087		855-97-0	3',4',5,7-Tetramethoxyflavone	I	I	Train
5-088		973-67-1	5,6,7-Trimethoxyflavone	I	I	Val
5-089		218-01-9	Chrysene	I	I	Train
5-090		238-84-6	Benzo[a]fluorene	I	I	Train

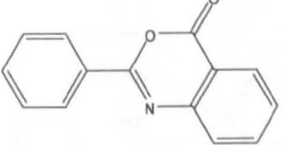
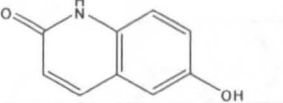
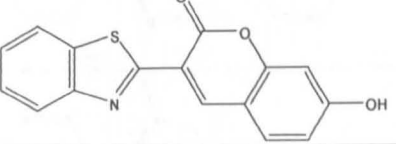
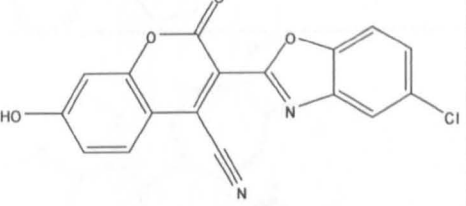
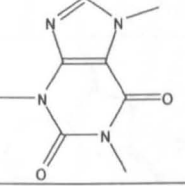
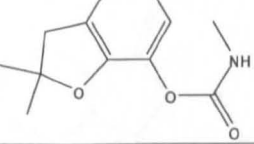
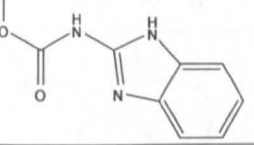
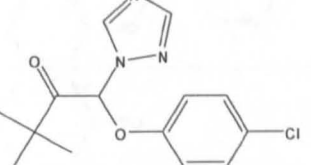
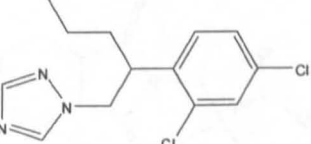
5-092		66-76-2	Dicoumarin	I	I	Train
5-093		6948-88-5	alpha-Naphtholbenzoin	A	A	Val
5-094		90-33-5	4-Methyl-7-hydroxycoumarin	I	I	Train
5-095		485-72-3	Formononetin	A	A	Train
5-097		4709-68-6	Benzhydrylidenefluorene	I	I	Val
5-098		1801-42-9	2,3-Diphenyl-1-indenone	A	A	Val
5-099		596-01-0	alpha-Naphtholphthalin	A	A	Train

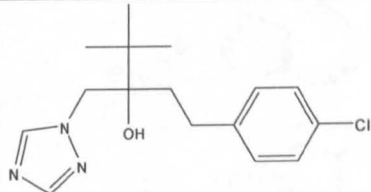
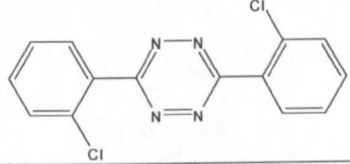
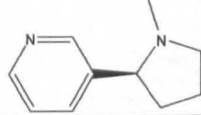
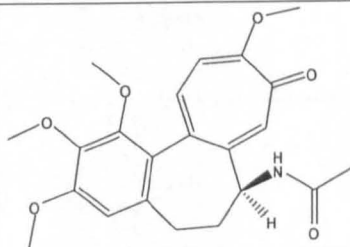
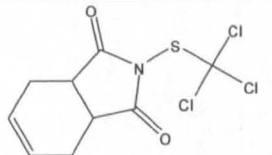
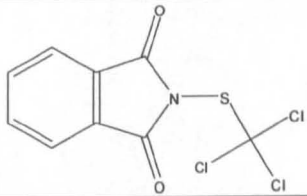
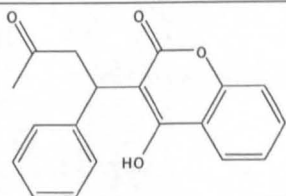
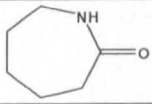
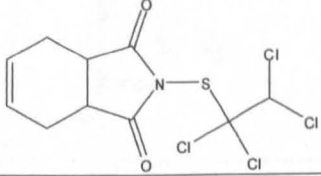

5-101		15499-84-0	4,4'-(9-Fluorenylidene)dianiline	I	I	Train
6-001		17924-92-4	Zearalenone	A	A	Train
6-002		42422-68-4	beta-Zearalanol	A	A	Train
6-004		71030-11-0	beta-Zearalenol	A	A	Val
6-005		5975-78-0	Zearalanone	A	A	Test
6-006		479-13-0	Coumestrol	A	A	Val
6-007		446-72-0	Genistein	A	A	Test

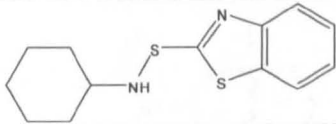
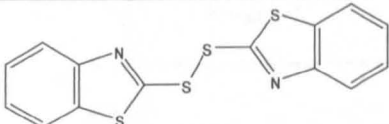
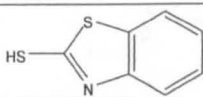
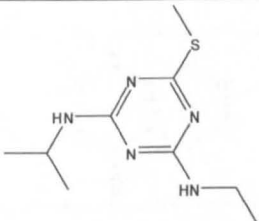
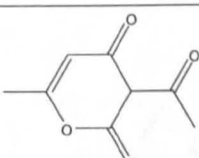
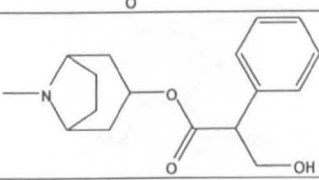
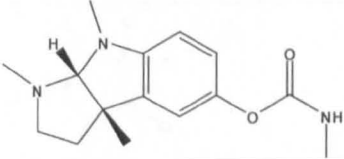

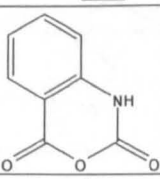
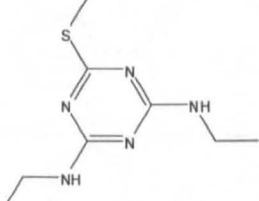
6-008		1912-24-9	Atrazine	I	I	Train
6-010		59-31-4	2-(1H)-Quinolinone	I	I	Test
6-011		15776-59-7	3-(2-(Diethylamino)ethyl)-7-hydroxy-4-methylcoumarin hydrochloride	I	I	Train
6-012		6093-71-6	7-Hydroxycoumarin-3-carboxylic acid ethyl ester	I	A	Train
6-013		6468-96-8	3-Phenylumbelliferone	A	A	Train
6-014		32809-16-8	Procymidon	I	I	Train
6-015		36734-19-7	Iprodione	I	I	Test
6-016		50471-44-8	Vinclozolin	I	A	Train
6-017		58-89-9	Lindane	I	I	Train

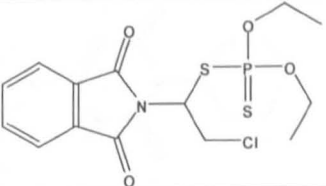
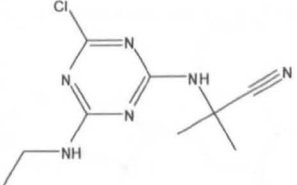
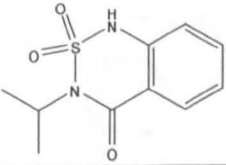
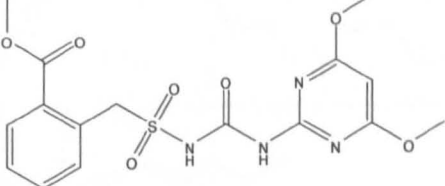
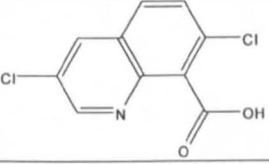
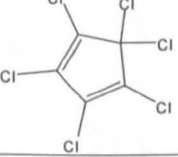
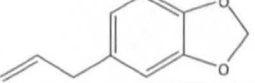
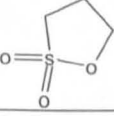
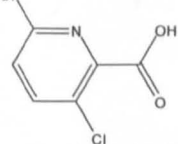
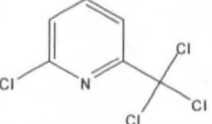
6-018		2385-85-5	Mirex	I	I	Train
6-019		143-50-0	Kepone	A	A	Train
6-022		52-86-8	Haloperidol	I	I	Train
6-023		60207-90-1	Propiconazole	I	I	Train
6-024		50-55-5	Reserpine	I	I	Train
6-025		60168-88-9	Fenarimol	A	I	Train
6-026		61-82-5	Amitrol = Aminotriazol	I	I	Train
6-027		50-06-6	Phenobarbital	I	I	Train
6-028		115-29-7	Endosulfan	A	I	Train

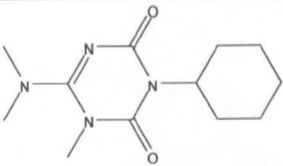
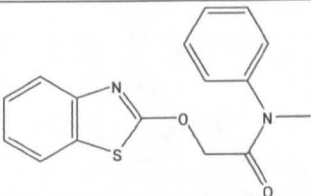
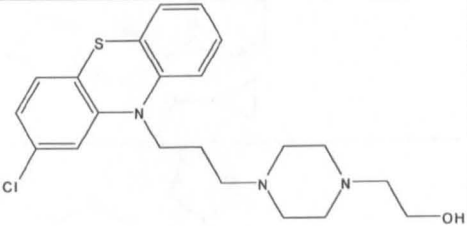
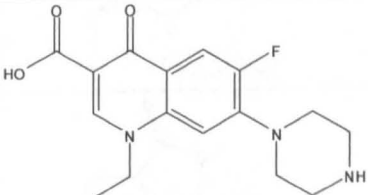
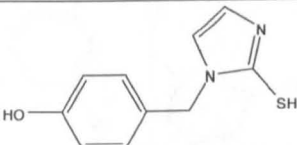
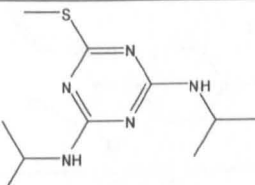
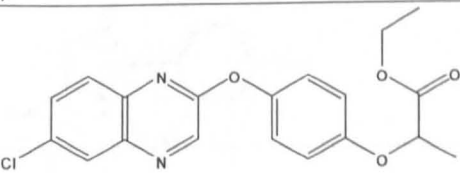
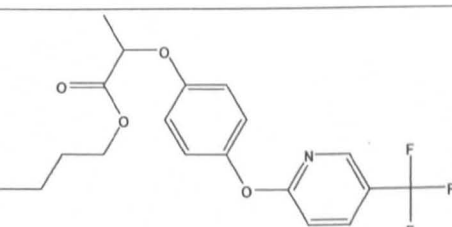
6-029		122931-48-0	Rimsulfuron	I	I	Train
6-030		114369-43-6	Fenbuconazole	I	I	Test
6-032		17804-35-2	Benomyl	I	I	Train
6-033		16069-36-6	Dicyclohexyl-18-crown-6	I	I	Train
6-035		2212-67-1	Molinate	I	I	Train
6-036		1214-39-7	N-6-Benzyladenine	I	I	Val
6-037		96-45-7	Ethylene thiourea	I	I	Test
6-038		612-96-4	2-Phenylquinoline	I	I	Train
6-039		14548-48-2	4-(4-Chlorobenzoyl)pyridine	I	I	Test

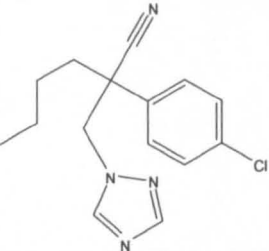
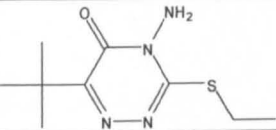
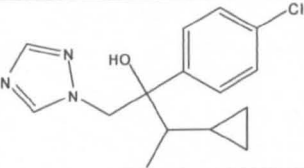
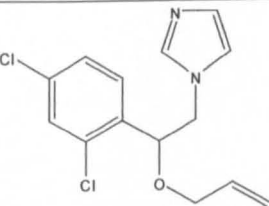
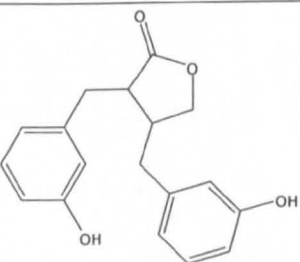
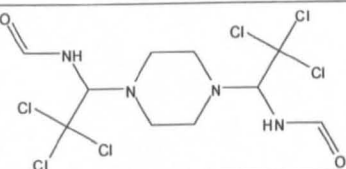
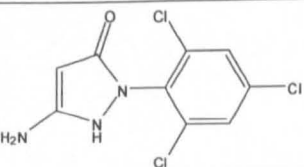
6-040		1022-46-4	2-Phenyl-4h-3,1-benzoxazin-4-one	I	I	Train
6-041		19315-93-6	2,6-Quinolinediol	I	I	Train
6-043		58851-99-3	3-(2-Benzothiazolyl)umbelliferone	A	A	Train
6-045		97477-81-1	3-(5-Chloro-2-benzoxazolyl)-4-cyanoumbelliferone	A	I	Train
6-046		58-08-2	Caffeine	I	I	Train
6-047		1563-66-2	Carbofuran	I	I	Train
6-048		10605-21-7	Carbendazim	I	I	Train
6-050		43121-43-3	Triadimefon	I	I	Val
6-053		66246-88-6	Penconazole	I	I	Train

6-054		107534-96-3	Tebuconazole	I	I	Train
6-056		74115-24-5	Clofentezine	I	I	Train
6-060		54-11-5	Nicotine	I	I	Train
6-061		64-86-8	Colchicine	I	I	Test
6-062		133-06-2	Captan (ISO) ; 1,2,3,6-tetrahydro-N-(trichloromethylthio)phthalimide	A	I	Train
6-063		133-07-3	N-(Trichloromethylthio)phthalimide	A	I	Test
6-065		81-81-2	Warfarin	I	I	Train
6-066		105-60-2	Caprolactam	I	I	Train
6-067		2425-06-1	Captafol (ISO); 1,2,3,6-tetrahydro-N-(1,1,2,2-tetrachloroethylthio)phthalimide	A	I	Val
6-068		24602-86-6	Tridemorph (ISO); 2,6-dimethyl-4-tridecylmorpholine	I	I	Train

6-071		95-33-0	N-Cyclohexyl-2-benzothiazolesulfenamide	A	I	Test
6-072		120-78-5	2,2'-Dithiobis[benzothiazole]	A	I	Train
6-074		149-30-4	2-Benzothiazolethiol	A	I	Train
6-075		834-12-8	Ametryn (ISO); 2-ethylamino-4-isopropylamino-6-methylthio-1,3,5-triazine	I	I	Train
6-076		16807-48-0	Dehydroacetic acid	I	I	Train
6-077		51-55-8	Atropine	I	I	Train
6-078		57-47-6	Physostigmine	I	I	Test
6-079		108-93-0	Cyclohexanol	I	I	Val
6-080		118-48-9	Isatoic anhydride	I	I	Test
6-081		1014-70-6	Simetryn (ISO); 2,4-bis(ethylamino)-6-methylthio-1,3,5-triazine	I	I	Test

6-082		10311-84-9	Dialifos (ISO) ; 2-chloro-1- phthalimidoeth yl O,O-diethyl phosphorodithi oate	A	I	Train
6-083		21725-46-2	Cyanazine	I	I	Train
6-084		25057-89-0	Bentazone (ISO); 3- isopropyl- 2,1,3- benzothiadiaz ine-4-one-2,2- dioxide	I	I	Train
6-085		83055-99-6	Methyl alpha- (4,6- dimethoxypyri midin-2- yl)ureidosulph onyl)-o-toluate	I	I	Train
6-086		84087-01-4	3,7- Dichloroquinoli ne-8- carboxylic acid	I	I	Train
6-087		77-47-4	Hexachlorocyc lopentadiene	A	I	Train
6-088		94-59-7	Safrole	I	I	Val
6-089		1120-71-4	1,2- Oxathiolane 2,2-dioxide	I	I	Val
6-090		1702-17-6	Clopyralid	I	I	Train
6-091		1929-82-4	2-Chloro-6- (trichlorometh yl)pyridine	I	I	Train

6-093		51235-04-2	3-Cyclohexyl-6-dimethylamino-1-methyl-1,2,3,4-tetrahydro-1,3,5-triazine-2,4-dione; hexazinone	I	I	Test
6-094		73250-68-7	2-(Benzothiazol-2-yloxy)-N-methyl-N-phenylacetamide; mefenacet	I	I	Train
6-095		58-39-9	4-(3-(2-Chlorophenothiazin-10-yl)propyl)-1-piperazineethanol	I	I	Train
6-096		70458-96-7	Norfloxacin	I	I	Train
6-098		95333-64-5	3-(4-Hydroxybenzyl)-imidazole-2-thione	I	I	Train
6-099		7287-19-6	Prometryn	I	I	Val
6-100		76578-14-8	Quizalofop-ethyl	I	I	Train
6-101		69806-50-4	Fluazifop-butyl	I	I	Val

6-102		88671-89-0	Myclobutanil	I	I	Test
6-103		64529-56-2	Ethiozin	I	I	Test
6-105		94361-06-5	Cyproconazole	I	I	Train
6-106		35554-44-0	Imazalil	I	I	Train
6-108		78473-71-9	Enterolactone	A	I	Train
6-110		26644-46-2	Triforine	I	I	Train
6-113		27241-31-2	3-Amino-1-(2,4,6-trichlorophenyl)-5-pyrazolone	A	I	Train